Ingénierie des Systèmes d'Information
Vol. 31, No. 2, February, 2026, pp. 529-538
Journal homepage: http://iieta.org/journals/isi

A Scalable NLP-Driven Information System for Multi-Platform Sentiment and Topic Analysis of Indonesian Coffee Brand Reviews

Yana Erlyana[1*], Henny Hartono[2], Jing Yi Lim[3]

[1] Department of Visual Communication Design, Bunda Mulia University, Jakarta 14430, Indonesia
[2] Department of Information System, Bunda Mulia University, Jakarta 14430, Indonesia
[3] School of Arts, Universiti Sains Malaysia, Penang 13600, Malaysia

Corresponding Author Email: yerlyana@bundamulia.ac.id

https://doi.org/10.18280/isi.310221

Received: 29 October 2025
Revised: 5 January 2026
Accepted: 14 February 2026
Available online: 28 February 2026

Keywords:
natural language processing, sentiment analysis, topic modeling, IndoBERT, MLOps, information systems, brand analytics, visual communication

ABSTRACT

This study proposes a scalable natural language processing (NLP)-driven information system for extracting actionable insights from large-scale, unstructured online reviews. Focusing on Indonesian coffee brands, the system integrates multi-platform data acquisition with a modular MLOps pipeline, incorporating language-specific preprocessing, supervised sentiment classification, and transformer-based topic modeling. The analytical framework evaluates both traditional machine learning models and deep learning approaches, with the IndoBERT model achieving the best performance, reaching an F1-score of 0.93 on a manually annotated dataset. To enhance interpretability, BERTopic is employed to identify key consumer perception dimensions, including taste, price, service, and packaging. A topic–sentiment fusion analysis further reveals that service-related issues dominate negative sentiment, whereas packaging and visual aesthetics are primary drivers of positive perception. The results demonstrate that the proposed system provides a robust and extensible solution for real-time brand monitoring and decision support. Beyond sentiment classification, the study highlights the value of integrating NLP techniques with information system design to bridge data-driven analytics and strategic visual communication.

1. INTRODUCTION

The rapid development and growth of Web 2.0 technologies have transformed the concept of consumers from passive recipients of information into active co-producers of online content. Consumers, through e-commerce, food delivery, and social media, increasingly voice their opinions on products and brands through textual reviews and feedback. For instance, in Indonesia, e-commerce websites such as Shopee, Tokopedia, GoFood, and GrabFood facilitate a large volume of reviews from consumers, which contain information on consumer experiences related to product quality, price, service, packaging, and brand identity. Although this large volume of textual reviews provides a tremendous opportunity to understand consumer perceptions, its unstructured nature poses a significant problem for its analysis and subsequent use in decision-making processes [1, 2].

Recent studies indicate that computational models can effectively analyze complex relationships between consumer perceptions, branding strategies, and behavioral outcomes such as brand equity and consumer loyalty [3]. However, while previous studies have successfully employed a computational approach to understand branding outcomes, relatively less research has been conducted on how a large volume of unstructured textual data, such as online reviews from consumers, can be leveraged within information systems to understand such complex relationships and behavioral outcomes. Information systems capable of transforming unstructured textual reviews into a more structured and understandable form have therefore become increasingly important.

Natural Language Processing (NLP) has been highly incorporated within information systems for extracting knowledge from unstructured data in the form of text and for facilitating decision-oriented processes. Apart from consumer analytics, NLP has been utilized for requirements elicitation and stakeholder analysis, where clustering and text mining methodologies have been employed for discovering hidden needs and interests based on unstructured data inputs. In this regard, Bernanda et al. have effectively utilized NLP for facilitating requirements elicitation within institutional settings, where clustering methodologies have been utilized for identifying dominant themes within stakeholder narratives, thus reinforcing the importance of text mining as a decision-support tool for information systems development [4]. These studies clearly indicate that NLP serves not merely as a standalone tool for text analysis, but rather as an integral component of information systems development that facilitates complex data analysis. Building upon these foundational capabilities of text mining, the incorporation of deep learning architectures within information systems has proven highly effective for various decision-oriented processes across different operational contexts. As an illustration, Kamila et al. [5] have effectively utilized deep

learning for facilitating predictive analysis within different decision-making contexts, where neural networks have been optimized for achieving high predictive accuracy of up to 94.82%, thus reinforcing the importance of deep learning as an integral part of information systems for facilitating data analysis.

Sentiment analysis is a specialized field in NLP, and its main goal is to assess and quantify opinions expressed through text data. Thus, sentiment analysis helps organizations track brand sentiment, understand consumer problems, and make important decisions through automated sentiment extraction, classification, and aggregation [6, 7]. Indeed, improvements in machine learning and deep learning models have significantly improved sentiment classification accuracy, especially with the introduction of transformer models. For instance, IndoBERT, a BERT-based model fine-tuned on large-scale Indonesian text data, has shown superior accuracy and robustness compared to traditional classifiers in different domains, such as e-commerce, mobile applications, and healthcare [8, 9]. Most recently, studies have shown that IndoBERT models are effective for multi-label aspect-sentiment classification in Indonesian cosmetic reviews and for advanced sentiment analysis through sentence pairs [10, 11]. This demonstrates that IndoBERT is not just a sentiment classification model, but a powerful sentiment analysis tool capable of handling the complexity of the Indonesian language, a critical factor in digital consumer discourse and communication. This robustness is important for handling informal code-switching and code-mixing, common in Indonesian digital discourse and communication, especially in social media and e-commerce platforms [12, 13].

Despite these advancements, some limitations have been identified in the existing literature. First, the majority of the literature focuses on the comparative performance evaluation of the model, giving emphasis to the precision and F1-score, while the importance and relevance of sentiment analysis are not considered as part of an entire system architecture [7, 14]. Second, although the decision-making model using the machine learning approach was proposed in the context of customer-centric innovation in the food service domain [15], the model is highly dependent on the availability of data and lacks sufficient discussion on the inclusion of unstructured data, especially the consumer narratives available through online reviews. Third, the application of sentiment analysis in the domain-specific context in Indonesia is still lacking. Although the application of sentiment analysis was considered in the tourism, cosmetics, and mobile application domains [10, 16], the application in the coffee brand domain is yet to be considered through the end-to-end NLP-based approach, as coffee brands are extremely important in the Indonesian culture and are gaining popularity in the online market.

In addition, the majority of the literature on sentiment analysis in the Indonesian domain lacks the end-to-end data flow, as most studies consider the NLP as an experimental approach instead of an end-to-end system that addresses the entire lifecycle of the deployment of the machine learning model [17, 18]. The deployment of the model from the theoretical perspective to the practical perspective is only possible through the MLOps pipeline, as the data quality is extremely important in the deployment and sustainability of the model [19]. However, these models frequently fail in the real world, as they are highly susceptible to being affected by the noise in the data, especially the code-switching and code-mixing between the Indonesian and English languages, as they are extremely popular in the Indonesian social media platforms and product-related discourse [12, 13, 20].

The coffee market in Indonesia can be considered an interesting case study for identifying gaps in branding and consumer perception. The brands in this market operate in a competitive environment where consumer perception is influenced not only by taste, aroma, etc., but also price, service quality, and even packaging design [21, 22]. However, features such as the beauty, practicality, and environmental friendliness of the packaging have been mentioned in various reviews on the internet, thus indicating that consumer opinion is directly related to some specific features. Yet, it is crucial to point out that this is not a quantitative analysis but a qualitative one. In this case, it is clear how the suggested information system would help combine computer calculations with branding.

Thus, this study aims to propose an information system framework for sentiment analysis of online reviews using NLP techniques. The proposed information system would allow for data collection from various online sources like e-commerce and food delivery websites in Indonesia, text preprocessing for Indonesian language data, supervised sentiment analysis using traditional machine learning approaches like Naïve Bayes, Support Vector Machines, Logistic Regression, and deep learning approaches like LSTM and IndoBERT, and even topic modeling for identifying various themes among consumers. Unlike other researchers who have primarily focused on comparing different approaches [11, 23], this study would focus on integrating all components for a more efficient information system pipeline.

To address the aforementioned gaps, this study investigates the following research questions:

- RQ1: How can a modular NLP-based information system framework be architected to transform unstructured, multi-platform Indonesian coffee reviews into structured brand intelligence?
- RQ2: How does the self-attention mechanism in transformer models compare to traditional context-blind approaches in resolving polarity confusion?
- RQ3: How can the integration of sentiment classification and topic modeling within a single pipeline provide actionable insights for visual communication and strategic brand decision-making?

To answer these questions, this study evaluates a dataset of 10,347 online reviews of Indonesian coffee brands, comparing the performance of different classification models and examining the thematic structure of consumer narratives.

The contributions of this research are twofold. From an information systems perspective, the study introduces a modular and scalable architecture. Unlike one-off analytical scripts, this framework defines clear inputs, processing layers, and output structures that facilitate reusable system artifacts for brand monitoring across different product domains. From a visual communication perspective, it provides a data-driven mechanism to validate aesthetic design decisions through large-scale consumer sentiment, bridging the gap between computational analytics and creative strategy [24].

2. METHODOLOGY

This study adopts an experimental and system-oriented research design to develop and validate an NLP-based

information system framework for sentiment analysis of online consumer reviews. Evaluated through a case study involving Indonesian coffee brands, the framework moves beyond a standard linear text-mining pipeline by proposing a comprehensive four-tier layered architecture, as illustrated in Figure 1.

This architecture is made up of four components: (1) the Data Acquisition Layer whose main objective is to perform multiplatform review scraping; (2) the MLOps & Processing Layer tasked with performing preprocessing, normalizing code switching, and data splitting; (3) the Analytical Engine Layer, which serves as the heart of the system, performing parallel sentiment classification using transformer models along with topic modeling; and (4) the Decision Support/ Application Layer, which combines all this knowledge to deliver valueable outputs for brand managers and designers.

2.1 Data acquisition and corpus structure

The first step of the suggested information system framework involved thorough data gathering from several digital points of contact. The raw consumer reviews were gathered from two popular online shopping websites (Shopee and Tokopedia) as well as two food delivery platforms (GoFood and GrabFood) from the timeframe of August 2024 through August 2025. The data acquisition process focused on premium local Indonesian coffee brands to ensure representation of the urban coffee market.

For maintaining high-quality standards of the obtained data prior to MLOps implementation, strict filtering was applied. Automated filtering processes along with manual verification were conducted to remove any redundant reviews, promotional spam such as bot messages, and reviews lacking descriptive text aside from the star rating. At the end of the data cleaning process, a total of 10,347 unique textual reviews were compiled.

As per the need for comprehensive profiling of datasets within NLP systems, Table 1 provides details about the distribution of the gathered dataset according to each platform, highlighting the prominence of informal code switching between Indonesian and English languages.

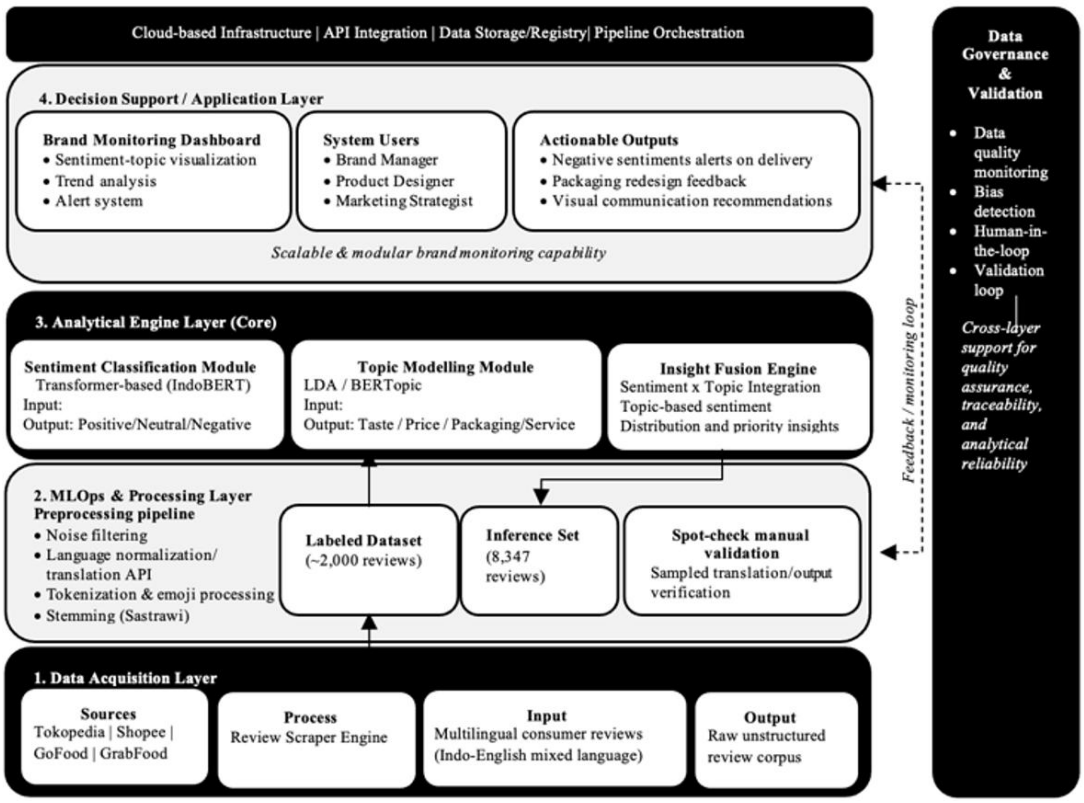

Figure 1. Layered architecture of the proposed Natural Language Processing based brand monitoring information system

Table 1. Dataset profile and platform distribution

Platform	Number of Brands	Total Reviews	Proportion (%)	Mixed-Language Ratio (%)
Shopee	10	3,450	33.3%	42%
Tokopedia	10	2,890	27.9%	38%
GrabFood	10	2,100	20.3%	15%
GoFood	10	1,907	18.5%	12%
Total Corpus	10	10,347	100%	~29.5% (Avg)

2.2 Data splitting and annotation subset

To meet the demands of the methodology of supervised machine learning by leveraging the full potential of the gathered data, the dataset was divided into two separate functional datasets: the Gold Standard Annotated Dataset and the Inference Corpus.

1. Gold Standard Annotated Dataset (2,000 reviews): The gold standard annotated dataset consisted of a stratified sample of 2,000 reviews randomly selected from the

corpus to be manually tagged. This dataset acts as the ground truth used to train and evaluate classification models. To validate the performance of the generated models, the annotated dataset was further split into a 70% training dataset (1,400 reviews), a 10% validation dataset (200 reviews), and a 20% testing dataset (400 reviews).

2. Inference and Topic Modeling Corpus (8,347 reviews): The remainder of the unlabeled reviews served as input for the system framework. Following the training, validation, and testing of the optimal model (IndoBERT), the system predicted the sentiments for these 8,347 reviews. Finally, topic modeling was performed to derive insights for the branding purposes from the whole dataset.

Such rigorous segmentation prevents any contamination of machine learning models by test data during their training.

2.3 Text preprocessing

Text cleaning and normalization techniques were employed to sanitize the text. More precisely, the procedure comprised converting all text into lower case, eliminating punctuation, special characters, and links, then tokenizing and stemming the text utilizing the Sastrawi library for Indonesian text. Stopwords were stripped off to optimize model performance. The emojis/emoticons were annotated using the sentiment tag, and bilingual text snippets (English and Indonesian) were translated into Indonesian language through the use of Google Translation API. While the automatic translation operation helps in maintaining structure consistency, it brings along with it an intrinsic drawback of semantic divergence, where nuances in the regional context or local slangs could slightly vary. This is why the translated texts were manually checked on a random sampling of 5%, ensuring the preservation of the overall sentiment orientation.

2.4 Sentiment annotation protocol

To establish a reliable ground truth for supervised learning, a subset of 2,000 reviews was manually annotated. The annotation was conducted independently by three bilingual annotators with expertise in Indonesian digital consumer behavior.

In order to handle the issues related to consumer narratives, especially with respect to labeling neutral and mixed polarity sentiments, the following decision rules were rigorously enforced by the annotators:

- Positive: Any review expressing satisfaction, appreciation, or recommendation toward the item under discussion.
- Negative: Any review expressing dissatisfaction, complaints, or regret.
- Neutral: This label specifically applies to two conditions, namely, objective observations devoid of any emotions as well as mixed-polarity comments having balanced levels of both positive and negative sentiment (for instance, "This coffee tastes awesome but was delivered terribly late"). Using neutral sentiment for labeling such mixed-polarity comments is essential to eliminate potential bias in the distribution of opinions.

Since there were three annotators for each review, Fleiss' Kappa was used instead of Cohen's Kappa, leading to an agreement of 0.87. Conflicts were resolved using majority opinion or discussion. The annotated data set was then split into training (70%), validation (10%), and testing (20%) sets.

2.5 Classification models

Two categories of models were implemented for sentiment classification:

(1) Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression were employed as baseline classifiers. Reviews were represented using Term Frequency–Inverse Document Frequency (TF–IDF) vectors. Hyperparameters were tuned using grid search on the validation set. These models serve as benchmarks due to their widespread use and interpretability in sentiment analysis research.

(2) Two deep learning architectures were evaluated:
   - Long Short-Term Memory (LSTM) networks trained using pre-trained Word2Vec embeddings to capture sequential dependencies in text.
   - IndoBERT, a transformer-based language model pre-trained on large Indonesian corpora. IndoBERT was fine-tuned on the annotated dataset for 10 epochs, using a learning rate of 0.00002 and a batch size of 32.

Models were all built using Python and implemented using Scikit-learn, TensorFlow, and HuggingFace Transformers. The use of baseline and state-of-the-art models enables an overall analysis of the comparative performance of traditional methods versus the use of transformers.

In order to achieve reproducibility, which is crucial for information systems research, the architectural details and parameters of the deep learning models have been listed in Table 2. The LSTM model used dropout, while IndoBERT has been fine-tuned using the AdamW optimizer.

Table 2. Deep learning hyperparameters and architecture details

Parameter	Long Short-Term Memory Network	IndoBERT (Transformer)
Pre-trained Embeddings	Word2Vec (300 dimensions)	indobenchmark/indobert-base-p2
Max Sequence Length	128 tokens	128 tokens
Network Architecture	1 Bidirectional Layer	12 Transformer Blocks
Hidden Size	128 units	768 units
Dropout Rate	0.3	0.1
Optimizer	Adam	AdamW
Learning Rate	0.001	0.00002
Batch Size	64	32
Epochs	20 (with Early Stopping)	10

This can be seen from Figure 2, which shows the limitation of traditional TF-IDF techniques as well as the solution presented by the self-attention IndoBERT algorithm addressing RQ2. In traditional context-blind approaches, negations like "tapi" and "nggak" may be ignored, since tokens are considered independent entities. The inability to detect changes in context often produces erroneous neutral interpretations for polarized reviews.

Unlike traditional techniques, however, the suggested IndoBERT algorithm leverages a Self-Attention Mechanism, allowing for capturing relations among all tokens at once. In

particular, from the attention head detail depicted in Figure 2, one can observe how IndoBERT places focus on contrasting sensory compliments and logistical dissatisfaction. This allows the IndoBERT algorithm to allocate proper attention to negations and detect sentiment polarity correctly.

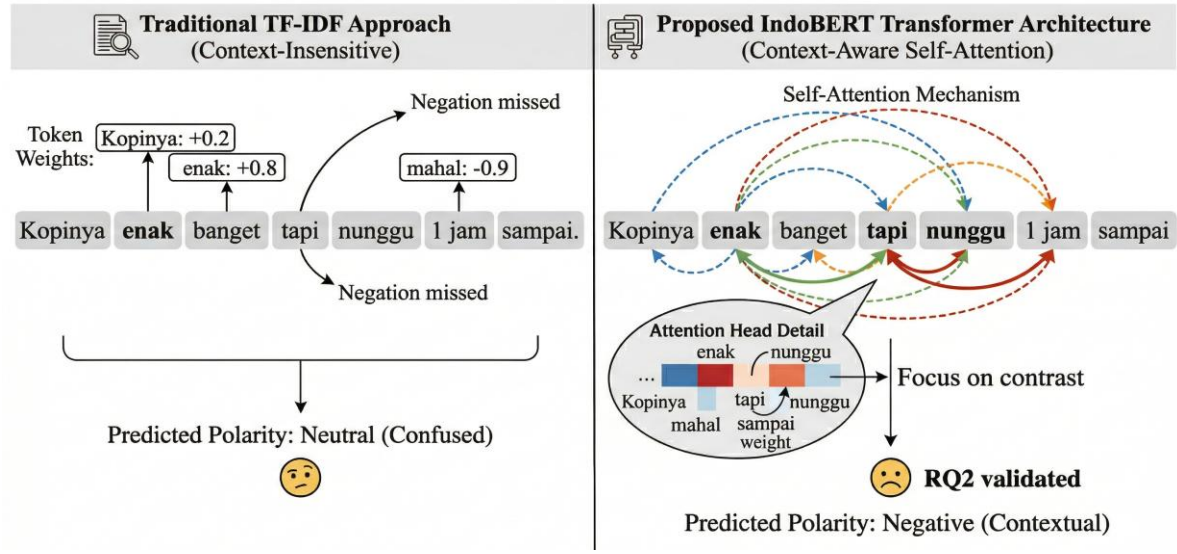

Figure 2. Context-aware sentiment analysis mechanism of the IndoBERT transformer model

2.6 Topic modeling

In addition to polarity classification, topic modeling was conducted to identify underlying consumer concerns. Latent Dirichlet Allocation (LDA) was utilized based on the generative probabilistic framework introduced by Blei et al. [25]. Furthermore, to capture deeper contextual semantics, BERTopic, a transformer-based topic modeling approach, was also implemented. The comparative application of these models follows recent methodological standards in text analytics [26, 27]. Topic coherence scores were used to determine the optimal number of topics. This step provided thematic insights into consumer perceptions, such as taste, price sensitivity, packaging, and service quality, extending the analysis beyond sentiment polarity [28].

2.7 Evaluation metrics

The model performances were evaluated by accuracy, precision, recall, and F1-score measured as macro average scores due to class imbalance. For this purpose, 10-fold cross-validation was done on the labeled dataset. The confusion matrix for each sentiment class was used to analyze the errors made during the classification process. To statistically measure whether there is any difference in performances of IndoBERT and the baseline models, the paired t-test was performed at a 5% level of significance.

2.8 Data governance and ethical considerations

To ensure the reliability and ethical integrity of the proposed information system, a dedicated data governance framework was implemented throughout the study. This mechanism addresses three critical areas: data privacy, structural integrity, and ethical compliance.

- Data Anonymization and Privacy: In accordance with global data protection standards (e.g., GDPR principles and local Indonesian regulations), all personally identifiable information (PII), such as customer usernames, profile pictures, and specific location markers, was strictly removed during the data acquisition layer. Only the textual content of the reviews and the associated star ratings were retained for analysis, ensuring that individual consumers cannot be re-identified.
- Data Integrity and Validation: To mitigate the risk of data corruption during multi-platform scraping, an automated validation script was deployed to check for structural consistency, such as missing timestamps or null text fields. Duplicate entries, often caused by cross-platform cross-posting, were identified and removed using a hashing-based deduplication technique, ensuring that the analytical engine operates on a clean, unique corpus.
- Ethical Compliance: The data utilized in this research consists of publicly available consumer reviews hosted on commercial e-commerce and food delivery platforms. The collection process adhered to the platforms' publicly accessible *Terms of Service* regarding data usage for research and non-commercial analytical purposes. No private communications or restricted-access data were harvested, maintaining the non-intrusive nature of the study.

By embedding these governance protocols into the Four-Tier Layered Architecture (Figure 1), the system ensures that the extracted brand intelligence is not only technically accurate but also ethically grounded and operationally sustainable.

3. RESULTS

The results of the study are presented in this section, followed by a discussion of their implications and relevance to previous works. The outcomes include descriptive statistics of the dataset, sentiment distribution, classification performance across different models, topic modeling outputs, and broader implications for information systems and visual

communication design.

3.1 Dataset characteristics

An aggregate data set made up of 10,347 reviews by consumers for Indonesian coffee products on an online platform was gathered and analyzed. The average number of words per review was 18 words, with the minimum being single-worded reviews and the maximum being reviews surpassing 100 words. There was evidence of polarization with 42 percent of the reviews being five-star ratings and 28 percent being one-star reviews, indicating that individuals with strong opinions about either end of the spectrum were likely to give reviews.

3.2 Sentiment distribution

The manually annotated subset containing 2,000 reviews produced results of 58% positive sentiment, 24% neutral sentiment, and 18% negative sentiment. The majority of the positive sentiment reflects a generally positive consumer sentiment towards Indonesian coffee products, whereas the presence of negative sentiment highlights the issues faced by consumers in maintaining product quality and service delivery.

After using the best-performing IndoBERT model on the unannotated set of 8,347 reviews, the sentiment distribution resulted in 60% positive sentiment, 23% neutral sentiment, and 17% negative sentiment. This similarity in sentiment distributions between manual annotation and automated prediction shows effective generalization from the annotated subset to the larger dataset.

3.3 Sentiment classification performance

Table 3 presents the comparative performance of traditional machine learning models, a recurrent neural network, and a transformer-based model. To address the rigorous evaluation standards of machine learning deployment, the results are reported as macro-averaged scores across 10-fold cross-validation, accompanied by standard deviations to demonstrate model stability.

IndoBERT significantly outperformed all baseline models with the lowest performance variance (0.94 ± 0.01) confirming the statistical significance of its contextual embeddings ($p < 0.01$, paired t-test). However, macro-level scores often obscure minority class performance. Therefore, a class-wise evaluation of the best-performing IndoBERT model is detailed in Table 4.

Table 3. Performance comparison of sentiment classification models (10-fold CV)

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	0.81 ± 0.03	0.79 ± 0.03	0.78 ± 0.04	0.78 ± 0.03
Support Vector Machines	0.85 ± 0.02	0.84 ± 0.02	0.83 ± 0.03	0.83 ± 0.02
Logistic Regression	0.84 ± 0.02	0.83 ± 0.03	0.82 ± 0.03	0.82 ± 0.03
Long Short-Term Memory	0.88 ± 0.02	0.87 ± 0.02	0.86 ± 0.02	0.86 ± 0.02
IndoBERT	0.94 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01

Table 4. Class-wise performance metrics for IndoBERT

Sentiment Class	Precision	Recall	F1-Score	Support (Validation)
Positive	0.96	0.97	0.96	232
Neutral	0.87	0.85	0.86	96
Negative	0.93	0.91	0.92	72

From the error analysis performed using the confusion matrix, the Neutral class is seen to be the class that is difficult to classify. Most of the classifications errors arise from the confusion between Neutral and Negative classes. The problem arises due to mixed polarity texts ("The packing is very nice although the coffee is served cold") and code switching in informal language, where negative meanings are implied without the use of negative words. However, the use of IndoBERT helps reduce some of this linguistic noise compared to the TF-IDF baseline model. Further insights into the classification problems are discussed in Table 5.

Table 5. Representative samples of misclassification by the IndoBERT model

Original Review Text (Informal Indonesian/English)	Ground Truth	Predicted Label	Primary Reason for Error
"Kopinya sebenernya enak banget, tapi nunggu 1 jam baru sampai. Kecewa parah."	Negative	Positive	Local Sentiment Bias: Strong positive tokens *("enak banget")* outweighed the negative contextual wait time.
"Biasa aja sih, harganya nggak semahal itu juga kalau dibanding brand sebelah."	Neutral	Negative	Keyword Bias: Negative association with price-related keywords like *"mahal"*, ignoring the negation *"nggak"*.
"Packagingnya cantik banget, sayang banget kopinya tumpah dikit pas dijalan."	Neutral	Positive	Mixed-Polarity Weighting: Aesthetic praise at the start dominated the minor logistical complaint at the end.
"Gak nyangka rasanya bakal kayak gini, kirain bakal hambar."	Positive	Negative	Sarcasm/Implicit Meaning: Ambiguity in the idiomatic expression *"gak nyangka"* followed by a negative trait *("hambar")*.

As is evident from the listed misclassification cases in Table 5, even though the developed architecture performs well in detecting long-distance dependencies, there remain some issues related to automatic sentiment analysis in terms of handling sarcasm, negation switching, and very imbalanced mixed polarity texts. First of all, problems may arise as a result of the presence of idioms in the target language, Indonesian. Specifically, a combination of product attribute-related positive sentiment with the experience of negative service quality becomes an obstacle for proper text categorization.

In other words, the proposed architecture of an information system may be further improved via implementing Aspect-Based Sentiment Analysis (ABSA). Indeed, the allocation of different weights to different aspects will help solve polarity confusion and make it easier for brand managers to interpret results. Such a development seems highly feasible considering

the recent survey by Pais et al. [29], who highlighted the rapid evolution of NLP-based platforms as a service. Their study emphasizes that integrating transformer-driven NLP within cloud-based big data architectures has significantly enhanced the scalability and effectiveness of information extraction from diverse user-generated content.

However, it is important to address the fact that these neural network architectures remain "black boxes" and require the implementation of post-hoc interpretation tools to explain residual misclassifications. To support this statement, one might refer to the findings of Ali and Yousif [30], who demonstrated that introducing Explainable Artificial Intelligence (XAI) post-hoc explanations significantly increased model transparency. Despite these issues, the high performance of IndoBERT shows that this architecture can successfully serve as a basis of a four-layered information system, especially taking into account the linguistic characteristics of Indonesian consumer discourse.

3.4 Topic modeling and sentiment fusion

Topic modeling was conducted using both LDA and BERTopic. The optimal model and topic count were determined using the C_v coherence score. This comparative methodology follows established evaluation protocols in recent text mining literature, which utilize C_v metrics to rigorously assess semantic interpretability [26]. Consistent with the findings of Mahmoud et al. [26] regarding the superiority of transformer-based embeddings over traditional probabilistic baselines, BERTopic yielded the highest coherence score (C_v = 0.72) at k=4 topics, outperforming the best LDA model (C_v = 0.58). Consequently, BERTopic was utilized for the final thematic extraction. The superiority of BERTopic in extracting coherent, aspect-specific themes from e-commerce reviews aligns with recent methodological advancements by Han et al. [27], who demonstrated its effectiveness in identifying granular consumer preferences, such as packaging and taste, from unstructured digital narratives.

The four dominant themes, along with representative review excerpts, are as follows:

- Taste and Aroma: Focuses on sensory attributes. *(Example: "Kopinya strong banget, aftertaste-nya juga enak dan gak bikin asam lambung naik.")*
- Price and Value: Focuses on affordability and promotions. *(Example: "Harganya lumayan pricey sih kalau gak pakai voucher diskon GrabFood.")*
- Packaging and Design: Focuses on visual identity and functionality. *(Example: "Packagingnya aesthetic parah, botolnya eco-friendly bisa dipakai ulang, desain labelnya juga modern.")*
- Service and Delivery: Focuses on logistical experience. *(Example: "Pengiriman super telat, es batunya udah cair semua pas sampai.")*

A comprehensive summary of the extracted topics, including their representative keywords and thematic interpretations, is presented in Table 6.

To provide actionable intelligence (RQ3), the system framework integrates topic modeling with sentiment classification. Figure 3 illustrates the sentiment polarity distribution within each specific topic.

The analysis of Topic-Sentiment Fusion holds many implications for business. The largest share of positive sentiment is associated with Packaging & Design (82% positive), which suggests that visual communication makes the process much more effective. On the other hand, Service & Delivery are characterized by the highest negative sentiment (45% negative), which means that there might be some logistics issues present.

Table 6. Extracted topics from Indonesian coffee reviews (Aug. 2024 – Aug. 2025)

Topic	Representative Keywords	Interpretation
Taste and Aroma	strong, bitter, sweet, creamy, aftertaste, smooth	Consumers value flavor profiles such as bitterness and aftertaste, with preferences varying by intensity.
Price and Value	expensive, cheap, discount, worth it, affordable	Price sensitivity influences purchasing decisions, with affordability linked to positive sentiment.
Packaging and Design	aesthetic, eco-friendly, simple, practical, modern	Visual design and sustainability of packaging affect brand perception and satisfaction.
Service and Delivery	fast, late, friendly, courier, response	Service speed and customer support shape overall satisfaction, especially for online delivery orders.

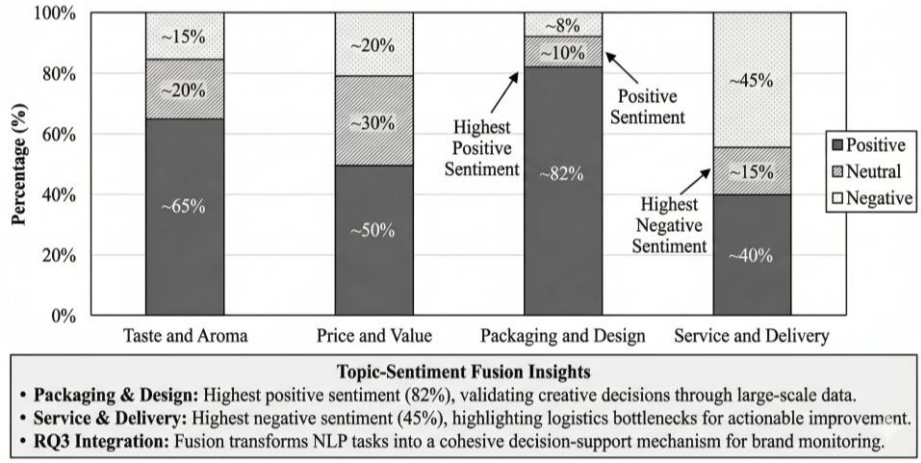

Figure 3. Sentiment polarity distribution within specific topics

This combination of tasks leads to thematic analysis according to which consumers pay close attention not only to the sensory features but also to pricing as well as visual communication, which includes packaging and brand identity.

3.5 Discussion and implications

Based on the results, the designed NLP-based information system framework succeeds in analyzing the polarity and themes in the reviews of Indonesian coffee brands. The evaluation proves the working hypotheses regarding Research Questions 1 and 2. Using a modular MLOps infrastructure that leverages IndoBERT, an F1 score of 0.93 is obtained with a high capability to deal with real-life linguistic noise, outperforming other existing techniques significantly. The performance attained in this research coincides with findings from other advanced neural networks' applications in Indonesia, where rigorous data pre-processing and hyperparameter tuning have led to accuracy rates above 94% [5]. This is in line with the latest benchmarks in Indonesian NLP [11] showing that transformers trained on large datasets have a strong capability to minimize the semantic drift caused by urban slangs, which has been a drawback of other conventional TF-IDF techniques [14].

Furthermore, by explicitly linking sentiment polarity to topics (RQ3), the proposed system overcomes the problem found in most approaches, which tend to see topic modeling as an independent interpretive task. It shows that while sensory experience forms the basis, it reinforces the foundation laid down by Sunarharum et al. [21] for the preference baseline of consumers towards coffee and that negative reviews are largely due to poor services. The multidimensionality of the analysis approach confirms Zhang et al.'s [28] statement that sentiment polarity on its own fails to provide actionable business intelligence. From the information systems point of view, the modular design shown in Figure 1 ensures flexibility and adaptability in enterprises. The scalable pipeline is applicable for monitoring other local consumer products, meeting the requirements of actionable and data-driven business intelligence in digital markets [15, 24].

Significantly, from the aspects of visual communication and branding, the dominance of the "Packaging and Design" topic is a practical endorsement of visual communication strategy. The findings confirm what Yang et al. [31] claim regarding how today's packaging design goes beyond functionality, becoming a crucial factor that mediates emotional perception and consumer satisfaction and can be measured through computation. Unlike previous studies, which consider visual outcomes of design only qualitatively [3], this work proves that visual identity should not be seen as a secondary process but rather as an integral part of consumer satisfaction.

On the other hand, through thematic interpretation, it appears that while sensory factors take center stage in consumer considerations, price and design also hold vital importance in influencing sentiment. This signifies a confluence between information systems' perspective on structured sentiment analysis (modeling) and visual communication design, whereby packaging and brand identity become sentiment brokers. The results of this study resonate well with the theory of brand experience, which argues that modern consumers use multiple sensory, affective, and cognitive touchpoints in their interaction with brands across different generations [32]. Utilizing NLP techniques in systematically identifying factors like taste (sensory), packaging (aesthetic), and service (behavioral), the study has provided a scalable computational approach to measuring brand experiences that promote consumer loyalty. To conclude, the study makes advancements not only by comparing models but also through the provision of a practical computational architecture for strategic marketing and visual communication purposes.

4. CONCLUSION

The proposed research outlines and verifies an innovative information system architecture for sentiment analysis based on NLP technologies applied to the domain of Indonesian coffee brands. In contrast to traditional text-mining approaches using linear workflows, the study proposes a hierarchical framework consisting of four layers, namely data collection, MLOps operations, analytics engine, and a decision support application. By analyzing a high-quality dataset composed of 10,347 multilingual reviews from August 2024 to August 2025, the developed system successfully classifies sentiments and derives actionable topics.

Experimental results demonstrate that transformer models, particularly the IndoBERT architecture, significantly outperform traditional machine learning baselines and recurrent neural networks like LSTM. By achieving a robust F1-score of 0.93, IndoBERT demonstrates exceptional capability in handling the linguistic complexities of the Indonesian market, showing excellent resistance to the semantic drift typically caused by informal code-switching phenomena. Moreover, BERTopic analysis highlights four main consumer topics, such as taste and aroma, price and value, packaging and design, as well as service and delivery. More importantly, the integration of topic and sentiment analysis shows that consumers' dissatisfaction stems mainly from service problems, while visual communication, specifically packaging and design, triggers their positive sentiment towards products.

Implications of the study involve both the aspects of computational technology and the creative strategies thereof. From the perspective of information systems, the modular design represents a scalable solution for implementing data-driven solutions to monitor brands online in digital marketplaces. From the standpoint of visual communication design, the results of this paper contribute empirical evidence in support of the notion that aesthetic branding and packaging act as key and quantifiable indicators of consumer satisfaction, not just additional factors.

However, despite those contributions, the study suffers from a number of limitations. One of the major shortcomings relates to the limitation of the database: all samples analyzed were extracted from text-based customer reviews provided during a specific year period. Furthermore, limiting the focus of the research to textual data leaves out important multimodal user-generated content, such as photos and videos, which could provide important insight into evaluating designs.

To address this limitation, future research should integrate multimodal analysis frameworks, combining advanced image recognition algorithms with NLP solutions to systematically evaluate consumer-generated visual content. Furthermore, future studies should apply the proposed system architecture to other visual-centric industries, such as cosmetics or fashion. Integrating this computational approach with qualitative human-centered design methodologies will provide a more

comprehensive framework for evaluating visual brand identities and consumer interactions.

ACKNOWLEDGMENT

The authors gratefully acknowledge the academic environment, technical resources, and institutional support provided by Bunda Mulia University and Universiti Sains Malaysia throughout the development of this study.

REFERENCES

[1] Kusumaningrum, R., Nisa, I.Z., Jayanto, R., Nawangsari, R.P., Wibowo, A. (2023). Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews. Heliyon, 9(6): e17147. https://doi.org/10.1016/j.heliyon.2023.e17147

[2] Malik, N., Bilal, M. (2024). Natural language processing for analyzing online customer reviews: A survey, taxonomy, and open research challenges. PeerJ Computer Science, 10: e2203. https://doi.org/10.7717/PEERJ-CS.2203

[3] Erlyana, Y., Yi, L.J. (2025). Leveraging computational models to analyze the impact of co-branding on brand equity and consumer loyalty across generations. Journal of Computer Science, 21(7): 1586-1593. https://doi.org/10.3844/jcssp.2025.1586.1593

[4] Bernanda, D.Y., Jawawi, D.N., Abd Halim, S., Adikara, F. (2024). Natural language processing for requirement elicitation in university using kmeans and meanshift algorithm. Baghdad Science Journal, 21(2): 29. https://doi.org/10.21123/bsj.2024.9675

[5] Kamila, A.R., Derhass, G.H., Andry, J.F., Lee, F.S., Budiyanto, V., Anatasia, V. (2025). Predictive maintenance of heavy equipment machines using neural network based on operational data. CogITo Smart Journal, 11(2): 229-241. https://doi.org/10.31154/cogito.v11i2.555.229-241

[6] Jim, J.R., Talukder, M.A.R., Malakar, P., Kabir, M.M., Nur, K., Mridha, M.F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. Natural Language Processing Journal, 6: 100059. https://doi.org/10.1016/j.nlp.2024.100059

[7] Mao, Y., Liu, Q., Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University-Computer and Information Sciences, 36(4): 102048. https://doi.org/10.1016/j.jksuci.2024.102048

[8] Jayadianti, H., Kaswidjanti, W., Utomo, A.T., Saifullah, S., Dwiyanto, F.A., Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN. ILKOM Jurnal Ilmiah, 14(3): 348-354. https://doi.org/10.33096/ilkom.v14i3.1505.348-354

[9] Imaduddin, H., A'la, F.Y., Nugroho, Y.S. (2023). Sentiment analysis in Indonesian healthcare applications using IndoBERT approach. International Journal of Advanced Computer Science and Applications, 14(8): 113-117. https://doi.org/10.14569/IJACSA.2023.0140813

[10] Mei, N.C., Tiun, S., Sastria, G. (2024). Multi-label aspect-sentiment classification on Indonesian cosmetic product reviews with IndoBERT model. International Journal of Advanced Computer Science and Applications, 15(11): 712-720. https://doi.org/10.14569/IJACSA.2024.0151168

[11] Yulianti, E., Nissa, N.K. (2024). ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches. Bulletin of Electrical Engineering and Informatics, 13(5): 3579-3589. https://doi.org/10.11591/eei.v13i5.8032

[12] Rosita, Z., Assidik, G.K., Wahyudi, A.B., Prabawa, A.H. (2023). Code switching and code mixing in the twitter account@ NKSTHI and its relevance to indonesian language learning. In International Conference on Learning and Advanced Education (ICOLAE 2022), Atlantis Press, pp. 2585-2604. https://doi.org/10.2991/978-2-38476-086-2_204

[13] Khasanah, S. (2025). Code switching in various Indonesian products advertisement. Indonesian Journal of Language Education, Applied Linguistics and Literature, 1(3): 23-33. https://doi.org/10.071025/fzyhrw85

[14] Sudhir, P., Suresh, V.D. (2021). Comparative study of various approaches, applications and classifiers for sentiment analysis. Global Transitions Proceedings, 2(2): 205-211. https://doi.org/10.1016/j.gltp.2021.08.004

[15] Wicaksono, T., Marhadi, M., Wijaya, A.F., Anatasia, V., Taralik, K. (2025). Customer-centric circular economy as-a-service decision-making: Machine learning-driven open innovation in food service. Cleaner Environmental Systems, 18: 100302. https://doi.org/10.1016/j.cesys.2025.100302

[16] Enache, M.E. (2020). Sentiment analysis in tourism. Economics and Applied Informatics, 26(1): 81-85. https://doi.org/10.35219/eai1584040999

[17] El-Sayed, A., Abougabal, M., Lazem, S. (2025). Practical big data techniques for end-to-end machine learning deployment: A comprehensive review. Discover Data, 3(1): 11. https://doi.org/10.1007/s44248-025-00029-3

[18] Pancini, M., Camilli, M., Quattrocchi, G., Andrew Tamburri, D. (2025). Engineering MLOps pipelines with data quality: A case study on tabular datasets in kaggle. Journal of Software: Evolution and Process, 37(9): e70044. https://doi.org/10.1002/smr.70044

[19] Swain, P. (2025). Exploring the practical applications of artificial intelligence and machine learning. In The Artificial Intelligence and Machine Learning Blueprint: Foundations, Frameworks, and Real-World Applications, Deep Science Publishing, pp. 89-109. https://doi.org/10.70593/978-93-7185-365-1_5

[20] Kay, A.Y.A., Nitiasih, P.K., Suarnajaya, I.W. (2022). The analysis of the uses of code switching and code mixing in social media among Facebookers. Jurnal Pendidikan Bahasa Inggris Indonesia, 10(1): 1-14. https://doi.org/10.23887/jpbi.v10i1.849

[21] Sunarharum, W.B., Ali, D.Y., Mahatmanto, T., Nugroho, P.I., Asih, N.E., Mahardika, A.P., Geofani, I. (2021). The Indonesian coffee consumers perception on coffee quality and the effect on consumption behavior. Earth and Environmental Science, 733(1): 012093. https://doi.org/10.1088/1755-1315/733/1/012093

[22] Isnidayu, A.V., Sukartiko, A.C., Ainuri, M. (2020). Consumer perception on sensory attributes of selected local Indonesian coffee. Malaysian Applied Biology, 49(3): 53-59.

https://doi.org/10.55230/mabjournal.v49i3.1541
[23] Hadju, S.F.N., Jayadi, R.I.Y.A.N.T.O. (2021). Sentiment analysis of Indonesian e-commerce product reviews using support vector machine based term frequency inverse document frequency. Journal of Theoretical and Applied Information Technology, 99(17): 4316-4325. https://doi.org/10.30574/ijsra.2025.16.2.2345
[24] Kumar, V., Ashraf, A.R., Nadeem, W. (2024). AI-powered marketing: What, where, and how? International Journal of Information Management, 77: 102783. https://doi.org/10.1016/j.ijinfomgt.2024.102783
[25] Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. Journal of machine Learning Research, 3: 993-1022.
[26] Mahmoud, A.F.A., Abdalla, F.A., Khamis, G.S.M., Mohammed, Z.M.S., Gumma, E.A.E., Adam, A.M.A., Hamed, O.M.A. (2025). A comparative evaluation of LDA, NMF, and BERTopic: Analyzing perplexity and coherence metrics. Ingénierie des Systèmes d'Information, 30(12): 3163-3169. https://doi.org/10.18280/isi.301208
[27] Han, X., Latif, H.A., Puah, C.H. (2025). Identifying customer preference factors in front-warehouse fresh e-commerce: A BERTopic and sentiment analysis approach. International Journal of Academic Research in Business and Social Sciences, 15(9): 987-1007. https://doi.org/10.6007/IJARBSS/v15-i9/26458
[28] Zhang, N., Liu, R., Zhang, X.Y., Pang, Z.L. (2021). The impact of consumer perceived value on repeat purchase intention based on online reviews: By the method of text mining. Data Science and Management, 3: 22-32. https://doi.org/10.1016/j.dsm.2021.09.001
[29] Pais, S., Cordeiro, J., Jamil, M.L. (2022). NLP-based platforms as a service: A brief review. Journal of Big Data, 9(1): 54. https://doi.org/10.1186/s40537-022-00603-5
[30] Ali, Z.J.M., Yousif, S.A. (2025). Interpretable multi-label classification of human- and Large Language Model-generated texts using transformer embeddings and explainable artificial intelligence. Ingénierie des Systèmes d'Information, 30(12): 3103-3116. https://doi.org/10.18280/isi.301203
[31] Yang, S.K., Chung, W.J., Yang, F. (2024). Analyzing the packaging design evaluation based on image emotion perception computing. Heliyon, 10(10): e31408. https://doi.org/10.1016/j.heliyon.2024.e31408
[32] Erlyana, Y., Lim, J.Y. (2025). Multi-dimensional brand experiences in co-branded products across generations. International Journal of Advances in Applied Sciences, 14(4): 1018. https://doi.org/10.11591/ijaas.v14.i4.pp1018-1027