# Implementation of Zero-Shot DeBERTa and IndoBERT for Aspect-Based Sentiment Analysis on Reviews of Five LLM Applications

Fabriyan Fandi Dwi Imaniawan[1*], Ragil Wijianto[2], Vadlya Maarif[3], Joko Dwi Mulyanto[4], Mustofa[5], Aprih Widayanto[6]
[1,2,3,4,5,6]Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, Kwitang, Kec. Senen, Kota Jakarta Pusat, DKI Jakarta, Indonesia
e-mail: [1]fabriyan.fbf@bsi.ac.id, [2]ragil.rgw@bsi.ac.id, [3]vadlya.vlr@bsi.ac.id, [4]joko.jdm@bsi.ac.id, [5]mustofa.muo@bsi.ac.id, [6]aprih.apz@bsi.ac.id

***Abstract*** *- Large Language Model (LLM) applications such as ChatGPT, Gemini, Copilot, Claude, and Perplexity have been massively adopted in Indonesia, yet user experience evaluation remains largely limited to global sentiment analysis. This study implements Aspect-Based Sentiment Analysis (ABSA) using a dual-Transformer approach: DeBERTa zero-shot for aspect extraction and IndoBERT for sentiment classification on 5,000 Indonesian-language reviews from the Google Play Store across four aspect categories. Manual validation by two annotators on 300 samples yielded Cohen's Kappa of* $\kappa = 0.59$ *(aspect) and* $\kappa = 0.44$ *(sentiment), both Moderate. Evaluation against the gold standard showed aspect accuracy of 37.5% (weighted F1 = 0.42) and sentiment accuracy of 64.7% (weighted F1 = 0.61). Sensitivity analysis across five hypothesis templates revealed inter-template Kappa of 0.19–0.63, confirming template selection impact on predictions. Comparative analysis reveals Copilot achieves the highest satisfaction (mean score 4.67), while Claude records the most complaints (36.9% negative). This study contributes a validated comparative ABSA framework for Indonesian-language LLM applications.*

***Keywords:*** *Aspect-Based Sentiment Analysis, Large Language Model, Zero-Shot Classification, IndoBERT, Google Play Store*

## INTRODUCTION

The development of the generative Artificial Intelligence (AI) ecosystem has brought a paradigmatic shift in human interaction with technology. Large Language Model (LLM) applications such as ChatGPT, Google Gemini, Microsoft Copilot, Claude, and Perplexity have now been widely adopted in Indonesia for various purposes, from writing assistance to conversational information retrieval. Bilal et al. (2025) confirm that ChatGPT, Copilot, and Gemini dominate user feedback on the Google Play Store with usage patterns that vary across platforms. A similar study by Navaratna & Saxena (2025) of 117,353 reviews from 10 AI chatbot applications shows that organic reviews on the Google Play Store are a rich source of public opinion data for evaluating real-world technology acceptance. This phenomenon is also supported by Hossain (2025), who demonstrated that early user feedback on generative AI applications significantly influences the sustainability of technology adoption.

Although user satisfaction evaluation through sentiment analysis has been widely conducted, most previous studies have two fundamental limitations. First, prior studies generally classify sentiment only at the document level without identifying the specific aspects that trigger the sentiment. For example, Sinaga et al. (2025) analyzed sentiment on ChatGPT, Gemini, and Copilot applications in Indonesia using IndoBERT with an accuracy of 96.38%; however, that approach remains a general sentiment approach and therefore cannot distinguish whether user dissatisfaction originates from answer quality, subscription pricing, or application interface. Second, most studies still rely on classical machine learning algorithms such as Support Vector Machine (SVM) and Naïve Bayes (NB). A comparative study by Srianan et al. (2025) on 505,980 reviews shows that Transformer models (RoBERTa) consistently outperform SVM and NB with an accuracy of 92.31%, especially on imbalanced datasets. Walji et al. (2025), through a systematic review, also confirm that NB tends to overfit majority classes, while SVM faces quadratic computational burdens on large-scale corpora.

Aspect-Based Sentiment Analysis (ABSA) offers a solution to the limitations of document-level sentiment analysis by decomposing a review into more granular aspect-sentiment pairs. Several Transformer-based ABSA studies have succeeded in specific domains: Harumy et al. (2024) applied BERT to ABSA on Halodoc application reviews with 83-95% accuracy, Zaid et al. (2025) used AraBERT for ABSA on tourism reviews with an F1-score of 0.97, and Liu et al. (2025) integrated syntactic structures with LLMs to improve ABSA performance.

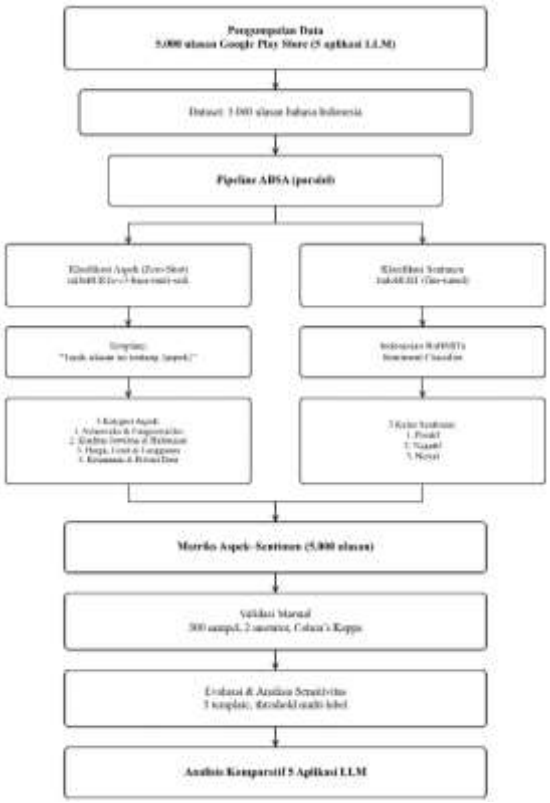
However, no study has specifically applied ABSA to the domain of generative AI applications. Bilal et al. (2025) analyzed user feedback on ChatGPT, Copilot, and Gemini, but it focused on use-case identification rather than aspect-level sentiment. Riccosan & Saputra (2025) applied multilabel sentiment analysis to Indonesian mobile application reviews using IndoBERT (F1 = 0.77), but did not explicitly implement aspect extraction. This gap is the main motivation for this study.

To address this gap, this study proposes a dual-Transformer approach that combines two pretrained models in a complementary manner. Multilingual DeBERTa-v3 is used for zero-shot aspect extraction based on Natural Language Inference (NLI), without requiring task-specific labeled data. The selection of DeBERTa is based on empirical evidence by Jian et al. (2025) showing its advantage as an ABSA backbone, with a 3.13% Macro-F1 improvement over the second-best model, as well as a study by Chattu et al. (2025) showing that DeBERTa- and XLM-RoBERTa-based architectures are effective for low-resource languages. For sentiment classification, the Indonesian RoBERTa Base Sentiment Classifier (IndoBERT) is used because it has proven reliable in recognizing sentiment from informal Indonesian text. Shaw et al. (2025) show that IndoBERT transfer learning is effective even with 1,000 data points, while Afuan et al. (2025) achieved 91% accuracy in Indonesian tweet sentiment analysis after fine-tuning IndoBERT.

Through a large-scale comparison of five leading LLM applications, this study aims to: (1) implement a dual-Transformer ABSA pipeline (zero-shot DeBERTa and IndoBERT) for aspect extraction and sentiment classification on 5,000 Google Play Store reviews; and (2) map the functional strengths and weaknesses of each application based on four evaluation aspect categories. The main contribution of this study lies in the first application of zero-shot DeBERTa-based ABSA combined with IndoBERT to analyze reviews of generative AI applications in Indonesia, an approach that has not been found in previous studies.

## RESEARCH METHOD

This study uses a quantitative text-mining approach with Transformer architecture to analyze sentiment in user reviews. The overall research workflow is visualized in Figure 1.



Source: Research results (2026)

Figure 1. ABSA Pipeline Workflow Using DeBERTa and IndoBERT

### 2.1. Data Collection

Review data were collected using a web scraping technique through the Python google-play-scraper library on the public Google Play Store interface. Similar approaches have been validated by Setiawan et al. (2024), who used the same module to extract 5,200 Tokopedia reviews and by Castilani & Tuga (2025), who applied Google Play scraping techniques to collect Traveloka application reviews. The data retrieval process focused on five LLM applications: ChatGPT, Gemini, Copilot, Claude, and Perplexity. Extraction parameters were limited to Indonesian-language reviews and sorted by the newest reviews. A total of 5,000 reviews were collected, with 1,000 reviews for each application. Descriptive statistics of the dataset are presented in Table 1.

Table 1. Descriptive Statistics of the Review Dataset for Five LLM Applications

| Application | Total | Mean Score | Median | Std. Dev. | Date Range |
|---|---|---|---|---|---|
| ChatGPT | 1.000 | 4.53 | 5.0 | 1.12 | 25 Feb-1 Mar 2026 |
| Gemini | 1.000 | 4.09 | 5.0 | 1.54 | 24 Feb-1 Mar 2026 |
| Copilot | 1.000 | 4.67 | 5.0 | 0.93 | 7 Sep 2025-1 Mar 2026 |
| Claude | 1.000 | 4.07 | 5.0 | 1.52 | 26 Sep 2024-1 |

| Application | Total | Mean Score | Median | Std. Dev. | Date Range |
|---|---|---|---|---|---|
| | | | | | Mar 2026 |
| Perplexity | 1.000 | 4.54 | 5.0 | 1.11 | 19 Nov 2025-1 Mar 2026 |
| **Total** | **5.000** | **4.38** | **5.0** | **1.28** | – |

Source: Research results (2026)

Table 1 shows that all applications have a median score of 5.0 with an overall mean of 4.38, indicating a tendency toward positive sentiment dominance. The data collection period varies: ChatGPT and Gemini have the narrowest time range (less than one week), while Claude has the longest range (approximately 18 months), indicating a relatively lower volume of Indonesian-language reviews.

### 2.2. Definition of Aspect Categories

To map user evaluations in a structured manner, this study defines four priority aspect categories derived from the ABSA literature in the mobile application domain (Chaudhary et al., 2025; Harumy et al., 2024) and from initial data exploration. The operational definition of each aspect is presented in Table 2.

Table 2. Operational Definitions of Aspect Categories

| No. | Aspect Category | Operational Definition |
|---|---|---|
| 1 | Interface, Error, and Functionality | User evaluation of UI navigation, responsiveness, bugs, crashes, and technical application features. |
| 2 | Answer Quality and Hallucination | User evaluation of the accuracy, relevance, and reliability of text/media outputs generated by the AI model. |
| 3 | Price, Limit, and Subscription | User evaluation of prompt-quota policies, premium subscription prices, and free-feature limitations. |
| 4 | Data Security and Privacy | User evaluation of login processes, account verification, data protection, and user privacy. |

Source: Research results (2026)

These four aspect categories were selected because they cover the main evaluation dimensions of generative AI applications: technical functionality, core service quality, business model, and user data security.

### 2.3. ABSA Modeling Pipeline

Review texts were directly fed into the Transformer pipeline without destructive manual preprocessing stages such as stemming or stopword removal. This decision is based on findings by Bustamin et al. (2025) that Transformer-based models can handle informal Indonesian text without requiring extensive normalization. All inference was executed on a GPU to accelerate processing. The modeling process consists of two stages:

#### 2.3.1. Aspect Classification (Zero-Shot)

Each review text was input into the mDeBERTa-v3-base-mnli-xnli model, which operates based on the Natural Language Inference (NLI) principle. The model calculates the probability of semantic relatedness (entailment score) between the review text (premise) and the hypothesis template for each aspect category. To accommodate reviews that may discuss more than one aspect simultaneously (Jian et al., 2025; Prado-Sánchez et al., 2025), inference was performed in multi-label mode (multi_label=True), so that each aspect receives an independent probability score (0-1). The aspect with the highest score was selected as the dominant label for the main distribution analysis. In addition, all multi-aspect scores were stored for multi-label analysis: at a given threshold, 96.4% of reviews were classified into more than one aspect (mean of 3.1 aspects per review), but the average gap between scores was only 0.14, indicating that the NLI scores were close and that a low threshold was insufficiently discriminative. At a higher threshold (Riccosan & Saputra, 2025), the multi-aspect proportion decreased to 63.4% (mean of 1.8 aspects), producing a more realistic distribution. The main analysis still used the dominant single-aspect label because: (a) no multi-label gold standard was available for evaluation, and (b) the argmax approach provides the most conservative and replicable interpretation.

To test prediction robustness against linguistic variation, five hypothesis templates were evaluated:
T1: "This review topic is about {aspect}"
T2: "This review discusses {aspect}"
T3: "This review is related to {aspect}"
T4: "This user comment concerns {aspect}"
T5: "This feedback is associated with {aspect}"

Template T1 was selected as the main template based on zero-shot NLI literature conventions (Lashyn et al., 2025). This zero-shot approach eliminates the need for task-specific labeled data, thereby significantly reducing manual annotation cost and bias.

#### 2.3.2. Sentiment Classification

In parallel, each review text was processed by the Indonesian RoBERTa Base Sentiment Classifier (Afuan et al., 2025; Riccosan & Saputra, 2025) to predict sentiment polarity into three classes: Positive, Negative, or Neutral. This model is an IndoBERT variant that has been specifically fine-tuned on an Indonesian sentiment corpus, giving it strong capability to recognize informal expressions, slang, and abbreviations typical of Indonesian users (Bustamin et al., 2025; Hidayatullah et al., 2025).

The results of the two stages were combined to form an Aspect-Sentiment matrix, in which each review has one dominant aspect label and one sentiment label that serve as the basis for comparative analysis across applications.

### 2.4. Manual Validation and Model Evaluation

To measure the reliability of the ABSA pipeline, manual validation was conducted through human annotation. A total of 300 review samples were selected in a stratified manner from 5,000 data points based on the proportions of applications (60 per application), aspects, and sentiments. Two independent annotators who had received aspect-definition guidelines (Table 2) performed blind annotation without knowing the model predictions. Each annotator assigned one aspect label and one sentiment label to each review.

Inter-annotator reliability was measured using Cohen's Kappa with the following interpretation: Almost Perfect, 0.61-0.80 (Substantial), and 0.41-0.60 (Moderate). For reviews on which both annotators agreed, the label was directly used as the gold standard; for disagreements, resolution was conducted through discussion until consensus was reached. Model performance was then evaluated against the gold standard using Accuracy, Precision, Recall, and weighted F1-score metrics.

## RESULTS AND DISCUSSION

### 3.1. Model Performance Evaluation

The manual annotation results show that inter-annotator Cohen's Kappa for aspect classification was in the Moderate category with 85.3% agreement, while sentiment classification was also Moderate with 69.3% agreement. These values indicate that aspect classification was more consistent across annotators than sentiment classification, likely due to differences in perceived polarity in mixed-sentiment reviews.

Table 3. Aspect Classification Evaluation (mDeBERTa Zero-Shot)

| Aspect Category | Precision | Recall | F1 | Support |
|---|---|---|---|---|
| Interface & Functionality | 0.86 | 0.30 | 0.44 | 235 |
| Answer Quality & Hallucination | 0.29 | 0.70 | 0.42 | 47 |
| Price, Limit & Subscription | 0.10 | 0.69 | 0.17 | 13 |
| Data Security & Privacy | 0.00 | 0.00 | 0.00 | 4 |
| *Macro avg* | 0.31 | 0.42 | 0.26 | 299 |
| *Weighted avg* | 0.73 | 0.37 | 0.42 | 299 |

Source: Research results (2026)

Aspect accuracy was 37.5%, with weighted F1 = 0.42 and macro F1 = 0.26. The large difference between weighted and macro F1 is caused by a highly imbalanced gold-standard distribution: the "Interface" class dominates 78.6% of samples (235 of 299), while "Price" accounts for only 4.3% (13) and "Security" for 1.3% (4). This imbalance explains why macro F1 is very low; poor performance on minority classes with very small support receives the same weight. The mDeBERTa model shows high precision for "Interface" (0.86) but low recall (0.30), indicating that the model distributes predictions to other aspects, while human annotators are more conservative by classifying most reviews into a single dominant aspect.

Although aspect F1 at the individual level is low, distribution analysis of 5,000 reviews remains meaningful at the aggregate level for three reasons. First, model classification errors are systematic (among semantically adjacent aspects), not random, so relative proportion patterns across applications are maintained. Second, the Chi-square test on full versus temporal-overlap sentiment distributions shows consistency across all applications, confirming the stability of the global pattern. Third, the analysis of five templates shows that although individual labels may change, the ranking of dominant aspects across applications remains relatively stable for four of the five templates.

Table 4 presents a comparison of aspect distributions between model predictions and the gold standard on 300 validation samples. The mDeBERTa model distributes predictions more evenly across the four aspects, whereas human annotators classify 78.3% of reviews as "Interface." This distributional difference is consistent with the behavior of zero-shot NLI, which evaluates each aspect independently, unlike annotators who tend to select one most dominant aspect. Nevertheless, both distributions agree that "Security" is the rarest aspect, confirming that the model correctly captures the relative scarcity of the minority class.

Table 4. Comparison of Aspect Distribution in 300 Validation Samples

| Aspect | Model ($n$) | Model (%) | Gold ($n$) | Gold (%) |
|---|---|---|---|---|
| Interface | 81 | 27.0 | 235 | 78.3 |
| Quality | 112 | 37.3 | 48 | 16.0 |
| Price | 90 | 30.0 | 13 | 4.3 |
| Security | 16 | 5.3 | 4 | 1.3 |

Source: Research results (2026)

For sentiment classification, accuracy reached 64.7% (weighted F1 = 0.61). IndoBERT performed best on the Positive class (F1 = 0.84) but struggled with the Neutral class (F1 = 0.27) because recall was very low (0.17). Per-application evaluation shows that ChatGPT achieved the highest aspect F1 (0.48) and Perplexity achieved the best sentiment F1 (0.68), whereas Copilot was consistently the lowest (aspect 0.33; sentiment 0.51).

### 3.2. Template Sensitivity Analysis

Testing the five hypothesis templates on 300 samples produced significant variation, as presented in Table 5.

Table 5. Sensitivity Analysis of Five Hypothesis Templates

| ID | Template | Conf. | $\bar{\kappa}$ |
|---|---|---|---|
| T1 | "This review topic is about {}" | 0.46 | 0.50 |

| ID | Template | Conf. | $\bar{\kappa}$ |
|---|---|---|---|
| T2 | "This review discusses {}" | 0.45 | 0.55 |
| T3 | "This review is related to {}" | 0.46 | 0.53 |
| T4 | "This user comment concerns {}" | 0.60 | 0.37 |
| T5 | "This feedback is associated with {}" | 0.44 | 0.49 |

Source: Research results (2026)

Template T4 shows outlier behavior: the highest confidence (0.60) but the lowest mean Kappa (0.37), indicating that the phrase "This user comment concerns" drives the model toward more confident but less stable predictions. Conversely, T2 ("This review discusses") achieves the highest agreement, indicating the strongest consistency with other templates. Inter-template Kappa varies from 0.19 (T4 vs. T5) to 0.63 (T2 vs. T3), demonstrating that template selection is not a trivial decision and requires empirical validation. Template T1 was selected for the main analysis because it balances confidence and stability and aligns with literature conventions (Prado-Sánchez et al., 2025). Although T2 has a slightly higher agreement value (0.55 vs. 0.50), the difference of 0.05 is not practically significant and remains within normal inter-template variability.

### 3.3. Distribution of Aspect and Sentiment Classification

The ABSA pipeline categorized 5,000 reviews into an Aspect-Sentiment matrix. The global aspect distribution is presented in Table 6.

Table 6. Distribution of Aspect Classification in 5,000 Reviews

| No. | Aspect Category | Total | Percentage |
|---|---|---|---|
| 1 | Answer Quality and Hallucination | 1.886 | 37.7% |
| 2 | Price, Limit, and Subscription | 1.524 | 30.5% |
| 3 | Interface, Error, and Functionality | 1.355 | 27.1% |
| 4 | Data Security and Privacy | 229 | 4.6% |
| 5 | Others | 6 | 0.1% |
| | **Total** | **5.000** | **100%** |

Source: Research results (2026)

The "Answer Quality and Hallucination" aspect dominates 37.7% of reviews, indicating that Indonesian users pay the most attention to LLM output accuracy, in line with previous findings (Bilal et al., 2025). The "Price, Limit, and Subscription" aspect ranks second (30.5%), showing that monetization issues are a significant concern. The "Data Security and Privacy" aspect accounts for only 4.6%, suggesting that most users prioritize functional utility.

The global sentiment distribution is dominated by the Positive class (3,238 reviews, 64.8%), followed by Negative (1,382, 27.6%) and Neutral (380, 7.6%). The dominance of Positive sentiment indicates the high utility of the five applications, while the substantial proportion of Negative sentiment indicates the need for improvement. The low Neutral proportion indicates that users tend to express clear opinions (Hossain, 2025).

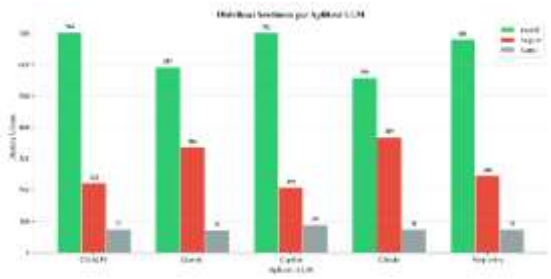
### 3.4. Comparative Analysis Across Applications

Table 7 presents the sentiment distribution per application, sorted by the highest proportion of positive sentiment.

Table 7. Sentiment Distribution per LLM Application

| Application | Positive | Negative | Neutral | Total | Mean Score |
|---|---|---|---|---|---|
| ChatGPT | 704 (70.4%) | 222 (22.2%) | 74 (7.4%) | 1.000 | 4.53 |
| Copilot | 702 (70.2%) | 209 (20.9%) | 89 (8.9%) | 1.000 | 4.67 |
| Perplexity | 681 (68.1%) | 246 (24.6%) | 73 (7.3%) | 1.000 | 4.54 |
| Gemini | 593 (59.3%) | 336 (33.6%) | 71 (7.1%) | 1.000 | 4.09 |
| Claude | 558 (55.8%) | 369 (36.9%) | 73 (7.3%) | 1.000 | 4.07 |

Source: Research results (2026)



Source: Research results (2026)
Figure 2. Comparison of Sentiment Distribution per LLM Application

Copilot ranks first (mean score 4.67; lowest negative proportion at 20.9%), although ChatGPT is slightly higher in absolute positive proportion (70.4%). In contrast, Claude and Gemini record the highest negative proportions (36.9% and 33.6%). Differences in use cases (Copilot for productivity versus ChatGPT for creative generation) (Bilal et al., 2025) may explain differences in user expectations and satisfaction.

The Chi-square test between sentiment distributions in the full period and the overlap period (25 February-1 March 2026) shows no significant difference across the five applications, indicating that temporal variation in data collection did not create systematic distributional bias. However, the monthly stability test for Perplexity shows significant fluctuation, so temporal interpretation for this application requires caution.

### 3.5. Aspect-Based Complaint Analysis (Negative Sentiment)

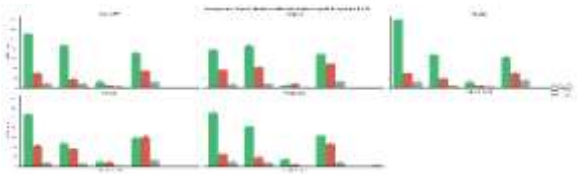
Analysis of 1,382 negative-sentiment reviews is presented in Table 8 and Figure 3.

Table 8. Distribution of Negative Sentiment by Aspect and Application

| Application | Interface | Price/ Limit | Security | Answer Quality | Total |
|---|---|---|---|---|---|
| ChatGPT | 43 | 90 | 13 | 76 | 222 |

| Application | Interface | Price/ Limit | Security | Answer Quality | Total |
|---|---|---|---|---|---|
| Copilot | 48 | 76 | 10 | 75 | 209 |
| Perplexity | 49 | 117 | 13 | 67 | 246 |
| Gemini | 104 | 123 | 18 | 91 | 336 |
| Claude | 94 | 151 | 20 | 104 | 369 |
| **Total** | **338** | **557** | **74** | **413** | 1.382 |

Source: Research results (2026)



Source: Research results (2026)

Figure 3. Comparative Aspect-Based Sentiment Analysis of Five LLM Applications

Several key findings emerge. First, the "Price, Limit, and Subscription" aspect is the largest source of complaints (557, 40.3%), with Claude being the highest (151) due to strict quota policies, confirming Navaratna & Saxena (2025). Second, "Answer Quality and Hallucination" ranks second (413, 29.9%), with Claude (104) and Gemini (91) recording the highest complaints due to hallucination phenomena (Bilal et al., 2025). Third, "Interface and Functionality" shows significant variation: Gemini dominates this category (104 complaints, 30.9% of its total complaints), mainly related to application crashes. Fourth, "Security" is the lowest (74, 5.4%), indicating that users prioritize utility over security.

### 3.6. Advantages of the Transformer Approach for ABSA

IndoBERT can classify sentiment from informal sentences containing slang and abbreviations typical of Indonesian users without requiring manual text normalization, consistent with Bustamin et al. (2025). Several comparative studies strengthen the justification for Transformer architecture: Aravind et al. (2025) showed that BERT outperforms SVM and NB on large-scale datasets, Srianan et al. (2025) reported that RoBERTa achieved 92.31% accuracy on 505,980 reviews (vs. SVM at 87.42%), and Hashmi & Yildirim (2025) showed that Transformer combinations achieved 94% accuracy.

The main advantage of the zero-shot DeBERTa approach is the elimination of labeled data requirements for aspect extraction. Unlike studies by Al-Dossari & Altalasi (2025) and Situmeang et al. (2025), which require labeled datasets, the NLI approach relies only on hypothesis templates for classification. However, the low aspect accuracy (37.5%) against the gold standard reveals that this approach still has significant limitations, especially in distinguishing the dominant aspect in reviews containing multi-aspect nuances. Chaudhary et al. (2025) reported zero-shot performance on application reviews with F1 = 0.842, but in a different domain and language. These findings emphasize the need for template and threshold calibration specific to each language domain.

## CONCLUSION

This study implements Aspect-Based Sentiment Analysis (ABSA) using zero-shot DeBERTa and IndoBERT on 5,000 reviews of five LLM applications on the Google Play Store, complemented by manual validation and sensitivity analysis. Based on the analysis, four main findings were obtained:

1. The "Answer Quality and Hallucination" aspect dominates user attention (37.7%), followed by monetization issues (30.5%) and interface functionality (27.1%). Copilot and ChatGPT show the highest satisfaction, while Claude and Gemini require significant improvement.
2. Manual validation by two annotators produced Cohen's Kappa values for aspects and sentiment, both in the Moderate category. Evaluation against the gold standard shows adequate sentiment accuracy of 64.7% (F1 = 0.61), but aspect accuracy of 37.5% (F1 = 0.42) reveals a significant limitation of the zero-shot model in distinguishing dominant aspects in multi-topic reviews.
3. Sensitivity analysis across five hypothesis templates reveals substantial variation (inter-template Kappa = 0.19-0.63), with one template showing outlier behavior, confirming that template selection requires empirical validation.
4. The Transformer approach is proven capable of processing informal text without destructive preprocessing and eliminating the need for labeled data, but it must be calibrated for each language domain.

Future research is recommended to: (a) develop fine-tuned models on Indonesian ABSA corpora to improve aspect accuracy; (b) expand temporal coverage through longitudinal monitoring; (c) explore optimal thresholds for multi-label aspect assignment; and (d) replicate the study on alternative platforms (iOS App Store) for cross-platform generalizability.

## REFERENCES

Afuan, L., Hidayat, N., Hamdani, H., Ismanto, H., Purnama, B. C., & Ramdhani, D. I. (2025). Optimizing BERT Models with Fine-Tuning for Indonesian Twitter Sentiment Analysis. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *16*(2), 248–267. https://doi.org/10.58346/JOWUA.2025.I2.016

Al-Dossari, H. Z., & Altalasi, M. (2025). Multi-Aspect Sentiment Analysis of Arabic Café Reviews Using Machine and Deep Learning Approaches. *Mathematics*, *13*(24), 3895. https://doi.org/10.3390/math13243895

Aravind, S. S., Rohilla, M., Kumar, O., Raj, S., & Sonal, D. (2025). Comparative study of BERT vs. traditional machine learning models for sentiment analysis. *AIP Conference Proceedings*, *3343*(1). https://doi.org/10.1063/5.0292610

Bilal, A., Mirza, H. T., Khan, A. S., Ahmad, A., Hussain, I., & Ali, S. Z. (2025). Who dominates generative AI? Analyzing user feedback to identify common use cases and areas for improvement in ChatGPT, Copilot and Gemini. *Knowledge and Information Systems*, *67*(11), 10797–10831. https://doi.org/10.1007/s10115-025-02550-y

Bustamin, A., Prayogi, A. A., Siswanto, D., Rafrin, M., & Nurdin, A. (2025). Text normalization for Indonesian slang words in sentiment analysis development. *ICIC Express Letters, Part B: Applications*, *16*(2), 121–129. https://doi.org/10.24507/icicelb.16.02.121

Castilani, L. A., & Tuga, M. (2025). Customer Sentiment Analysis on OTA Platforms: Insights for Enhanced User Experience and Service Optimization. *Proceedings of the 4th International Conference on Electronics Representation and Algorithm*, 316–321. https://doi.org/10.1109/ICERA66156.2025.11087277

Chattu, K., Reddy, K. A. N., Veesam, S. B., Chirumamilla, P. S., Babu, V. D., Prakash, K., Bansal, S. K., Faruque, M. R. I., & Al-Mugren, K. S. (2025). Sentiment classification for Telugu using transformed based approaches on a multi-domain dataset. *Scientific Reports*, *15*(1), 8124. https://doi.org/10.1038/s41598-025-05703-9

Chaudhary, M., Jain, C., & Anish, P. R. (2025). Exploring Zero-Shot App Review Classification with ChatGPT: Challenges and Potential. *Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering*, 672–677. https://doi.org/10.1145/3756681.3757036

Harumy, T. H. F., Pauzi, & Arian. (2024). Sentiment Analysis of Halodoc Application Reviews Based on Service Quality Aspects Using BERT. *Lecture Notes in Networks and Systems*, *1089*, 252–259. https://doi.org/10.1007/978-3-031-67195-1_30

Hashmi, E., & Yildirim, Ş. (2025). A robust hybrid approach with product context-aware learning and explainable AI for sentiment analysis in Amazon user reviews. *Electronic Commerce Research*, *25*(6), 5139–5171. https://doi.org/10.1007/s10660-024-09896-5

Hidayatullah, A. F., Apong, R. A. A. H. M., Lai, D. T. C., & Qazi, A. (2025). Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer. *Social Network Analysis and Mining*, *15*(1), 89. https://doi.org/10.1007/s13278-025-01444-9

Hossain, M. S. (2025). Emotional drivers of sustainable AI adoption: A sentiment analysis of early user feedback on the DeepSeek app. *Sustainable Futures*, *10*, 100947. https://doi.org/10.1016/j.sftr.2025.100947

Jian, Z., Li, J., Wang, M., Yao, J., & Wu, Q. (2025). Aspect sentiment learning for Aspect-Level Sentiment Classification. *Neural Networks*, *191*, 107758. https://doi.org/10.1016/j.neunet.2025.107758

Lashyn, Y., Trofymchuk, O. M., Zabolotnyi, S., Voitko, O., & Seabra, E. A. R. (2025). Sentiment analysis of texts using recurrent neural networks of the transformer architecture. *Advanced Information Systems*, *9*(3), 91–101. https://doi.org/10.20998/2522-9052.2025.3.11

Liu, H., Xiong, K., Wu, S., Cao, P., Cheng, K., & Liu, X. (2025). Integrating multiple syntactic structures for enhanced aspect-based sentiment analysis. *Engineering Applications of Artificial Intelligence*, *158*, 111297. https://doi.org/10.1016/j.engappai.2025.111297

Navaratna, A. R., & Saxena, D. K. (2025). Digital Governance Through Self-Regulation: A user-developer perspective of AI chatbots. *Journal of Telecommunications and the Digital Economy*, *13*(3), 1–29. https://doi.org/10.18080/jtde.v13n3.1077

Prado-Sánchez, V. P., Domínguez-Díaz, A., de-Marcos, L. O., & Martínez-Herráiz, J. J. (2025). Zero-Shot Classification of Illicit Dark Web Content with Commercial LLMs: A Comparative Study on Accuracy, Human Consistency, and Inter-Model Agreement. *Electronics*, *14*(20), 4101. https://doi.org/10.3390/electronics14204101

Riccosan, & Saputra, K. E. (2025). Multilabel classification sentiment analysis on Indonesian mobile app reviews. *IAES International Journal of Artificial Intelligence*, *14*(5), 4226–4234. https://doi.org/10.11591/ijai.v14.i5.pp4226-4234

Setiawan, I. H., Rahardi, M., Aminuddin, A., & Abdulloh, F. F. (2024). Sentiment Analysis of Tokopedia Application Reviews on Google Play Store Using BERT. *2024 International Conference on Information Technology Systems and Innovation*, 242–247. https://doi.org/10.1109/ICITSI65188.2024.10929357

Shaw, C., LaCasse, P. M., & Champagne, L. E. (2025). Exploring emotion classification of Indonesian tweets using large scale transfer learning via IndoBERT. *Social Network Analysis and Mining*, *15*(1), 67. https://doi.org/10.1007/s13278-025-01439-6

Sinaga, F. M., Pangaribuan, J. J., Kelvin, Ferawaty, & Widjaja, A. E. (2025). Dynamic Sentiment Analysis on the Emergence of Pre-Trained Generative Model-Based Applications in Indonesia. *International Journal of Advanced Computer Science and Applications*, *16*(12), 1150–1161. https://doi.org/10.14569/IJACSA.2025.01612111

Situmeang, S. I. G., Tambunan, S. R., Jevania, Simanjuntak, M. F., & Sinaga, S. (2025). Transformer and text augmentation for tourism aspect-based sentiment analysis. *IAES International Journal of Artificial Intelligence*, *14*(6), 4614–4622. https://doi.org/10.11591/ijai.v14.i6.pp4614-4622

Srianan, S., Nanthaamornphong, A., & Phucharoen, C. (2025). Advancing tourism sentiment analysis: a comparative evaluation of traditional machine learning, deep learning, and transformer models on imbalanced datasets. *Information Technology and*

*Tourism*, *27*(4), 1011–1045. https://doi.org/10.1007/s40558-025-00336-0

Walji, K., Erraissi, A., Zakrani, A., & Banane, M. (2025). From Review to Practice: A Comparative Study and Decision-Support Framework for Sentiment Classification Models. *International Journal of Advanced Computer Science and Applications*, *16*(9), 699–709. https://doi.org/10.14569/IJACSA.2025.0160967

Zaid, S., Alharbi, A., & Samra, H. El. (2025). Multi-Aspect Sentiment Classification of Arabic Tourism Reviews Using BERT and Classical Machine Learning. *Data*, *10*(11), 168. https://doi.org/10.3390/data10110168