Vol. 8, No. 4, August 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Rule-Based Aspect Extraction and IndoBERT-Based Sentiment Classification of Ruparupa Mobile Application Reviews

Erina Setyawati[1], Berlilana[2], Dhanar Intan Surya Saputra[3]

[1,2,3]Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia

Received:
February 4, 2026
Revised:
July 19, 2026
Accepted:
August 7, 2026
Published:
August 30, 2026

Corresponding Author:
Author Name*:
Erina Setyawati
Email*:
erinasetyawati92@gmail.com

DOI:
10.63158/journalisi.v8i4.1816

Abstract. This study evaluates a pipeline that separates rule-based aspect extraction from IndoBERT-based binary sentiment classification for Indonesian Ruparupa mobile application reviews. Google Play reviews were collected on 31 July 2026, anonymized, deduplicated before aspect expansion, and cleaned by lowercasing, removing URLs, emails, and special characters, and normalizing whitespace; no slang normalization, stop-word removal, or stemming was applied. Ratings 1-2 and 4-5 provided weak negative and positive labels, while three-star reviews were excluded. A 331-entry aspect dictionary mapped 1,495 unique reviews into 2,873 aspect-review pairs across six aspects. Across five repeated leakage-free group hold-out splits, IndoBERT achieved mean accuracy 0.9179 ± 0.0214, macro F1 0.9178 ± 0.0214, and ROC-AUC 0.9719 ± 0.0103; a calibrated TF-IDF + linear SVM baseline achieved 0.8765 ± 0.0124, 0.8759 ± 0.0127, and 0.9452 ± 0.0102, respectively. A McNemar test on run 1 showed a significant paired difference ($p = 0.00013$). Performance measures agreement with rating-derived weak labels rather than human-validated aspect sentiment. Because results from system-assigned aspects lacked independent human validation, aspect frequencies are descriptive rule-system outputs. Within this dataset, IndoBERT performed consistently across the five splits; supervised aspect extraction and human aspect-level annotation remain priorities.

Keywords: IndoBERT, Rule-Based Aspect Extraction, Weak Labeling, Google Play Review Mining, E-Commerce Reviews

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Mobile-application reviews document user reactions to software functionality, service failures, transactions, and post-purchase experiences. They therefore provide developers and service managers with direct, continuously updated user feedback that complements numerical ratings [1], [2]. In an e-commerce setting, one review may mention several touchpoints, including application usability, product availability, promotional mechanisms, payment, delivery, and complaint handling. Ruparupa was selected because it combines digital purchasing with omnichannel features such as Scan & Shop and store pickup across multiple home, hobby, and lifestyle retail brands, creating a review environment that covers both application and operational service issues.

Traditional document-level sentiment classification assigns one overall polarity to a review. Aspect-level sentiment classification instead predicts polarity toward a specified aspect, whereas aspect extraction determines which aspects occur in the text. These are distinct subtasks [3], [4]. SemEval formalized this separation through aspect-term or aspect-category identification and aspect-specific polarity prediction [5], [6]. A complete end-to-end ABSA system learns both subtasks, while a pipeline may combine a rule-based extractor with a supervised sentiment classifier.

Aspect extraction has been addressed through linguistic rules, unsupervised topic methods, supervised sequence labeling, multi-label classifiers, and transformer-based architectures [4], [7]. Rule-based extraction remains useful in low-resource settings because it is transparent and can be expanded with domain expressions, but it performs poorly on implicit aspects, polysemous keywords, sarcasm, and cross-aspect context. Recent e-commerce and Indonesian studies use supervised or deep models, including BiGRU-BiDAF, IndoBERT sentence-pair classification, CNN-based marketplace classification, and BERT-LSTM-CNN architectures [8], [9], [10], [11]. These studies show the value of aspect-level analysis while demonstrating the need for domain-specific annotation and careful preprocessing.

Transformer encoders learn contextual representations through self-attention [12]. BERT extends this architecture through bidirectional pre-training [13], while IndoNLU and IndoBERT provide Indonesian-language pre-training and evaluation resources [14], [15].

Contextual encoders are well suited to informal application reviews because a sentiment-bearing token can change meaning according to the aspect and surrounding words. Previous Indonesian mobile-review research has also reported strong BERT-based classification performance [16].

The study employed an expanded aspect dictionary containing 331 keywords and phrases, anonymized user identifiers, and removed duplicate reviews before multi-aspect expansion. Training, validation, and testing data were separated by review groups to prevent information leakage. Model performance was evaluated across five repeated group-based splits using class-wise, per-aspect, confidence-interval, and paired statistical analyses. The resulting framework combines rule-based aspect extraction with rating-derived binary sentiment classification rather than fully supervised end-to-end ABSA.

Binary polarity was used because one- and two-star ratings and four- and five-star ratings provide clearer weak anchors for negative and positive sentiment. Three-star reviews were excluded to reduce ambiguity in rating-derived supervision. This choice narrows the analyzed population because neutral ratings can contain mixed, moderate, or informational feedback about service experiences; the reported distributions therefore do not represent all Ruparupa reviews.

Accordingly, the research asks: (RQ1) What descriptive aspect and polarity patterns are produced by the rule-based aspect extraction system? (RQ2) How consistently does IndoBERT classify rating-derived sentiment across leakage-free group splits? (RQ3) How does IndoBERT compare with a calibrated TF-IDF + SVM baseline overall and by aspect? The contributions are a reproducible Indonesian e-commerce review pipeline, repeated group-based benchmarking with uncertainty estimates, and an explicit analysis of limitations created by weak sentiment labels and non-supervised aspect extraction.

2. METHODS

2.1. Research Design

The study used a modular computational text-analysis design. Figure 1 distinguishes the two core tasks: aspect assignment was performed using an expanded rule dictionary, while sentiment classification was conducted using IndoBERT and a classical machine-

learning baseline. All downstream evaluation used repeated group-based partitions so that aspect rows originating from the same review remained in one subset.

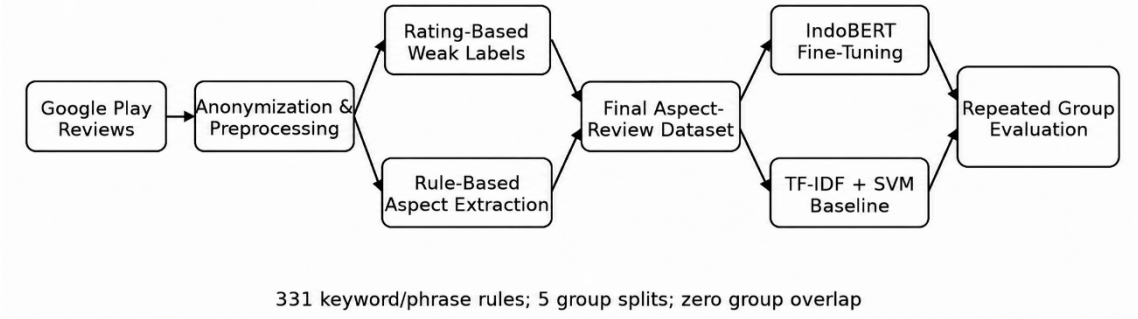

Figure 1. Research workflow for rule-based aspect extraction and sentiment classification

2.2. Data Collection

Reviews were retrieved from the Google Play Store with google-play-scraper version 1.2.7 using package com.mobileappruparupa, language id, country id, and newest-first sorting. The final scrape was executed on 31 July 2026. The raw file contained 3,222 reviews dated from 22 April 2019 to 29 July 2026. The source was publicly accessible; private accounts and restricted information were not accessed. User names and profile images were removed, and review identifiers were converted into SHA-256-based anonymous IDs before modeling. No usernames, profile images, or other identifiable reviewer information are reported in the manuscript or supplementary material.

Table 1. Data preparation and dataset summary

Stage	Procedure	Output
Raw collection	Google Play reviews; id-ID settings; newest-first	3,222 reviews
Cleaning and deduplication	Remove identifiers; normalize text; deduplicate by clean_review	1,938 unique cleaned reviews
Binary weak labels	Ratings 1-2 = negative; 4-5 = positive; rating 3 excluded	1,818 labeled reviews
Aspect expansion	Apply six rule-based aspect dictionaries to every review	2,873 aspect-review pairs

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Stage	Procedure	Output
Unique groups	Anonymous review IDs retained across expanded rows	1,495 review groups
Repeated evaluation	Five group splits; seeds 42, 52, 62, 72, 82	Zero group overlap in every run

2.3. Text Preprocessing and Weak Sentiment Labels

Preprocessing applied lowercasing; URL and email removal; removal of special characters; and whitespace normalization. The implemented clean_text() function did not use a slang-normalization dictionary, stop-word removal, stemming, or a dedicated typo-correction model. Because no stop-word filtering was applied, negation terms and aspect keywords remained in the cleaned review text. Deduplication occurred before aspect expansion to prevent repeated reviews from dominating the aspect-level dataset. Sentiment polarity was derived from review ratings. Ratings 1 and 2 were mapped to negative, ratings 4 and 5 to positive, and rating 3 was removed. This procedure is weak supervision: the rating describes overall satisfaction and may conflict with the written text or with sentiment toward an individual aspect [17]. Consequently, model scores quantify agreement with rating-derived labels rather than human-verified aspect-level sentiment.

2.4. Rule-Based Aspect Extraction

The aspect dictionary was developed through an error-driven examination of domain-specific expressions. The final resource contains 331 formal terms, informal expressions, and multiword phrases across six categories. Regex matching was applied directly to clean_review after basic cleaning. Each review was checked against every aspect dictionary. Because the pipeline did not apply stop-word removal or stemming, no protected-keyword selection step was required. Reviews matching several categories were expanded into multiple aspect-review pairs, while the first matching phrase within an aspect was stored only for traceability. Table 2 presents some Aspect definitions and representative Indonesian indicators.

Table 2. Aspect definitions and representative Indonesian indicators

Aspect	Operational Scope	Representative Indicators
App Quality	Application performance, login process, navigation, loading speed, application updates, checkout process	Slow performance, application crashes, login failure, checkout failure, loading issues, force close
Product Quality	Product condition, product suitability, stock availability, completeness, product availability	Damaged products, out-of-stock items, mismatched products, product defects, poor quality, authenticity issues
Price and Promotion	Product price, discounts, vouchers, cashback programs, promotional eligibility	Incorrect prices, discount issues, voucher cannot be used, cashback problems, promotions not applied
Delivery Service	Courier service, shipping cost, order tracking, delivery speed, packaging quality, arrival status	High shipping fees, courier issues, undelivered packages, delayed orders, late delivery, instant delivery problems, tracking issues
Customer Service	Complaint management, response quality, return process, refund service, customer support channels	Customer service issues, slow responses, complaints, return problems, refund difficulties, lack of assistance
Payment Process	Payment methods, bank transfer, payment verification, QRIS, paylater services	Payment failure, transfer problems, QRIS issues, paylater problems, verification delays, unsuccessful transactions

2.5. Group-Based Splitting and Model Configuration

The evaluation used five repeated group hold-out runs rather than k-fold cross-validation. In each run, GroupShuffleSplit used the review ID as the grouping variable. Each seed generated a different 72% training, 8% validation, and 20% testing partition. Verification showed zero group overlap among training, validation, and test subsets in all five runs. The validation subset selected the best IndoBERT epoch; the held-out test

subset was used only for final evaluation. This design reduces leakage caused by placing different aspect rows from one review in separate subsets [18], [19].

Table 3. Experimental configuration

Component	Configuration
IndoBERT	indobenchmark/indobert-base-p1; max length 128; batch 16; 3 epochs; AdamW; learning rate 2e-5; weight decay 0.01; dropout 0.1; linear warmup/decay; warmup ratio 0.1
Model selection	Validation macro F1 evaluated after each epoch; best checkpoint retained; no early-stopping patience rule
TF-IDF + SVM	Unigrams and bigrams; max_features 5,000; min_df 2; max_df 0.95; sublinear TF; LinearSVC C=1.0; balanced classes; CalibratedClassifierCV with sigmoid method and 3-fold calibration fitted within training only
Repeated evaluation	Five repeated group hold-out splits using seeds 42, 52, 62, 72, and 82; identical group-separated splits for both models; not k-fold cross-validation
Environment	Google Colab; Python 3.11; PyTorch 2.6.0; Transformers 4.51.3; scikit-learn 1.6.1; NVIDIA Tesla T4; CUDA 12.4

TF-IDF and SVM provide a transparent lexical baseline grounded in term weighting and maximum-margin classification [20], [21]. Scikit-learn implemented vectorization, calibration, metrics, and repeated splits [22]. Within each run, CalibratedClassifierCV (method='sigmoid', cv=3) fitted calibration folds only on the training subset; validation and test data were excluded from probability calibration.

2.6. Evaluation and Statistical Analysis

Each run reported accuracy; negative- and positive-class precision, recall, and F1; macro and weighted averages; ROC-AUC from actual positive-class probabilities; false positives; false negatives; and per-aspect metrics. Mean, sample standard deviation, and 95% t-based confidence intervals summarized five runs. McNemar's test with continuity correction compared paired predictions on run 1. Qualitative errors were inspected to identify rating-text disagreement, mixed-aspect reviews, implicit complaints, short expressions, and keyword ambiguity.

3. RESULTS AND DISCUSSION

3.1. System-Assigned Aspect and Sentiment Distribution

The expanded rule dictionary transformed 1,495 unique reviews into 2,873 aspect-review pairs. Product Quality was the most frequently assigned aspect (848; 29.52%), followed by App Quality (630; 21.93%) and Customer Service (499; 17.37%). Payment Process had the smallest support (161; 5.60%). These values describe the output of the rule system and are not presented as independently validated prevalence estimates.

Table 4. System-assigned aspect and rating-derived sentiment distribution

Aspect	Negative	Positive	Total	Share
App Quality	423	207	630	21.93%
Product Quality	329	519	848	29.52%
Price and Promotion	154	189	343	11.94%
Delivery Service	227	165	392	13.64%
Customer Service	230	269	499	17.37%
Payment Process	98	63	161	5.60%
Total	1461	1412	2873	100.00%

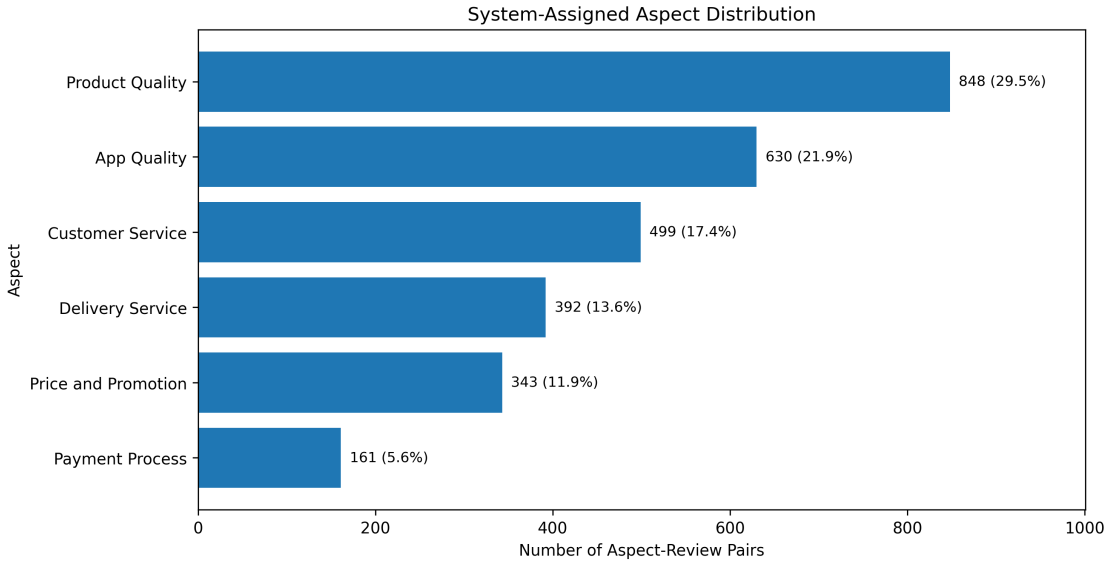

Figure 2. System-assigned aspect distribution in the final dataset

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

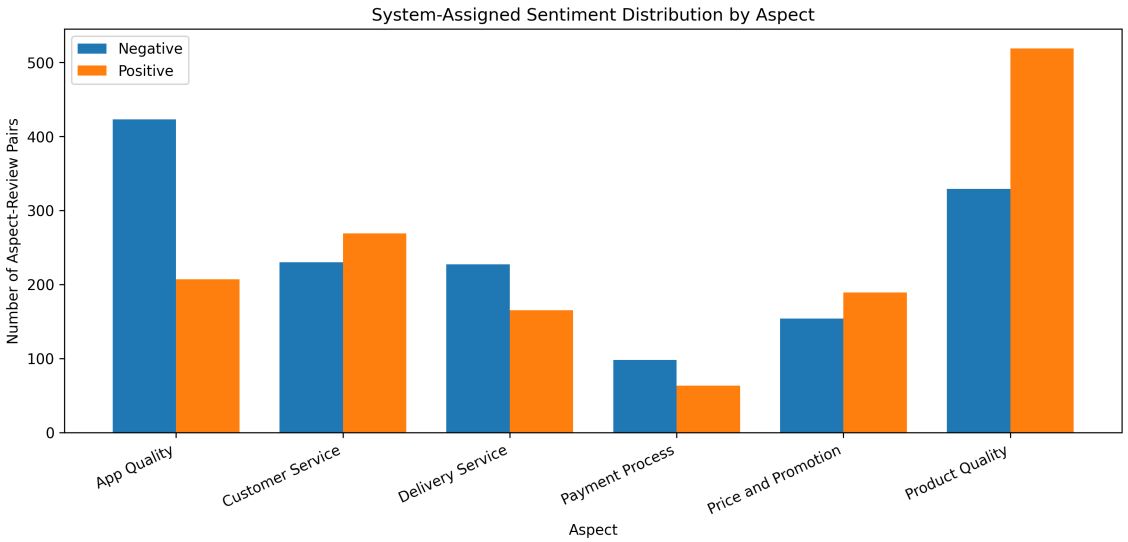

Figure 3. Rating-derived sentiment distribution by system-assigned aspect

3.2. Repeated Group-Based Model Performance

Across five repeated group hold-out splits in this dataset, IndoBERT achieved mean accuracy of 0.9179, macro F1 of 0.9178, and ROC-AUC of 0.9719. TF-IDF + SVM achieved 0.8765 accuracy, 0.8759 macro F1, and 0.9452 ROC-AUC. The confidence intervals in Table 5 show higher observed accuracy and macro F1 for IndoBERT in these splits, although five runs provide limited precision.

Table 5. Repeated group-based evaluation over five splits

Metric	IndoBERT mean ± SD	95% CI	TF-IDF + SVM mean ± SD	95% CI
Accuracy	0.9179 ± 0.0214	[0.8913, 0.9445]	0.8765 ± 0.0124	[0.8611, 0.8920]
Macro F1	0.9178 ± 0.0214	[0.8912, 0.9444]	0.8759 ± 0.0127	[0.8602, 0.8917]
Weighted F1	0.9180 ± 0.0213	[0.8915, 0.9445]	0.8762 ± 0.0127	[0.8605, 0.8919]
ROC-AUC	0.9719 ± 0.0103	[0.9592, 0.9847]	0.9452 ± 0.0102	[0.9324, 0.9579]

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

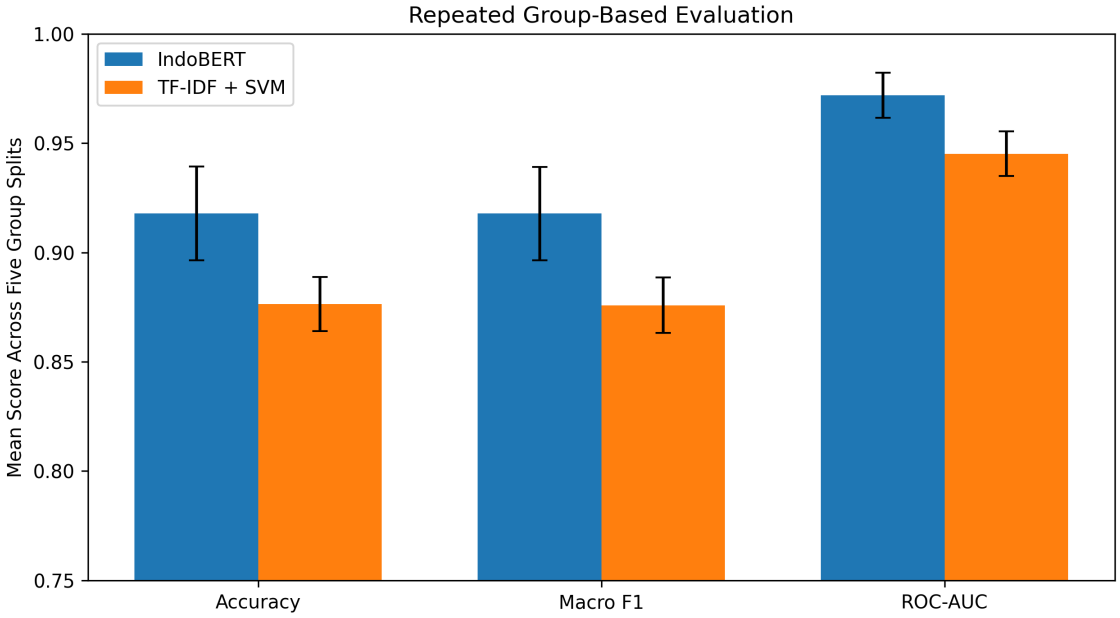

Figure 4. Mean performance across five group-based splits; error bars show standard deviation

3.3. Class-Wise Performance and Confusion Matrices

Run 1 used 580 held-out aspect-review pairs and is reported in detail for interpretability. IndoBERT classified 543 pairs correctly, compared with 508 for the baseline. Both confusion matrices contain 580 test instances and are consistent with the reported class-wise metrics.

Table 6. Class-wise metrics on run 1 (seed 42)

Model	Class/Average	Precision	Recall	F1	Support
IndoBERT	Negative	0.9416	0.9385	0.9400	309
IndoBERT	Positive	0.9301	0.9336	0.9319	271
IndoBERT	Macro Avg	0.9359	0.9360	0.9359	580
IndoBERT	Weighted Avg	0.9362	0.9362	0.9362	580
TF-IDF + SVM	Negative	0.8580	0.9191	0.8875	309
TF-IDF + SVM	Positive	0.8996	0.8266	0.8615	271
TF-IDF + SVM	Macro Avg	0.8788	0.8728	0.8745	580
TF-IDF + SVM	Weighted Avg	0.8774	0.8759	0.8754	580

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

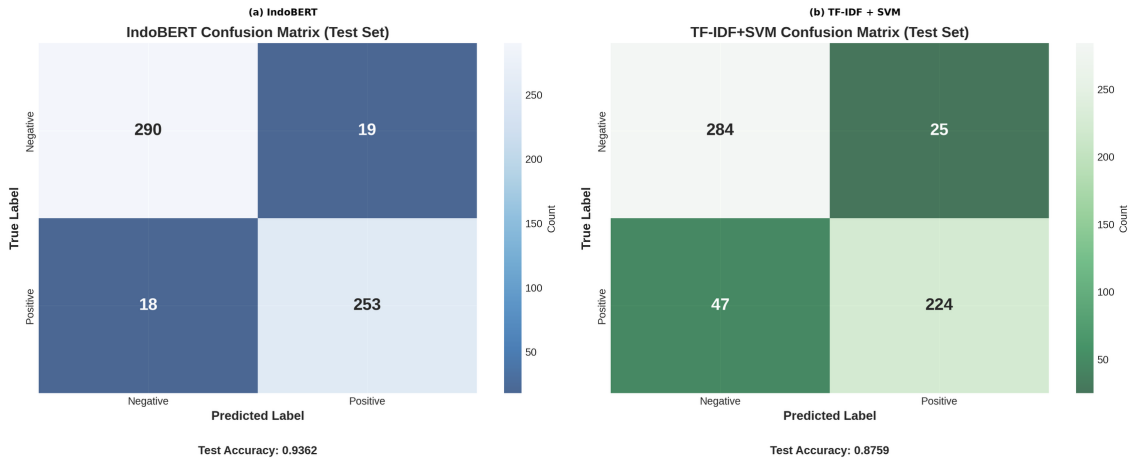

Figure 5. Run-1 confusion matrices: (a) IndoBERT and (b) TF-IDF + SVM

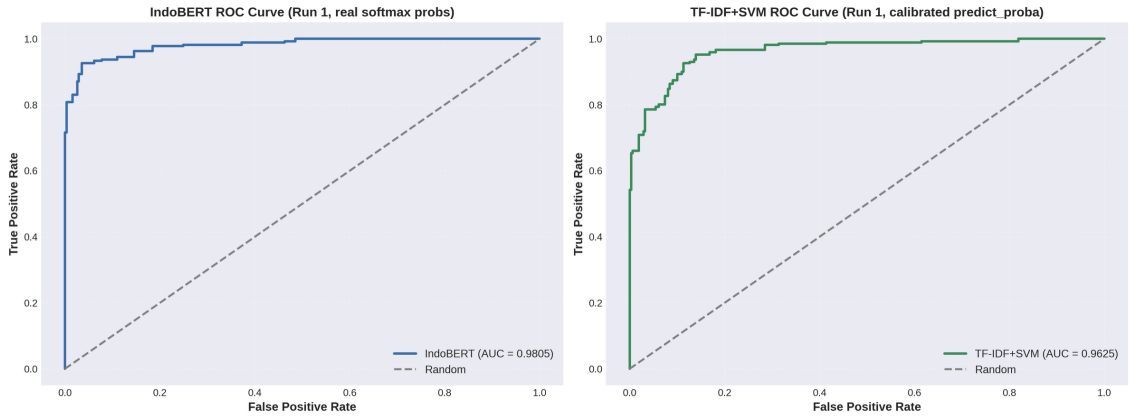

Figure 6. Run-1 ROC curves generated from actual model probabilities

3.4. Performance by System-Assigned Aspect

IndoBERT produced higher mean macro F1 for every system-assigned aspect across the five splits. Test-set support for each aspect is reported in run order in Table 7. Customer Service had the highest IndoBERT mean F1 (0.9456). Payment Process had the lowest mean F1 (0.8407) and the largest variation (SD 0.0789), consistent with its smaller support and more heterogeneous payment expressions. These per-aspect metrics evaluate rule-system assignments rather than human-validated aspect labels.

Table 7. Mean per-aspect macro F1 across five splits

System-assigned aspect (test support, run 1-5)	IndoBERT	TF-IDF + SVM	Relative improvement
App Quality (133/142/117/120/129)	0.8827 ± 0.0236	0.8434 ± 0.0229	4.66%
Product Quality (167/169/179/177/168)	0.9228 ± 0.0228	0.8754 ± 0.0117	5.41%
Price and Promotion (60/65/79/60/65)	0.9283 ± 0.0211	0.8988 ± 0.0273	3.28%
Delivery Service (82/77/84/76/66)	0.9065 ± 0.0105	0.8387 ± 0.0301	8.08%
Customer Service (107/111/82/99/94)	0.9456 ± 0.0160	0.9054 ± 0.0082	4.44%
Payment Process (31/41/37/27/24)	0.8407 ± 0.0789	0.7965 ± 0.0570	5.55%

3.5. Statistical Comparison and Error Analysis

On run 1, both models were correct on 486 pairs and both were wrong on 15. IndoBERT alone was correct on 57 pairs, while the SVM alone was correct on 22. McNemar's test yielded $\chi^2 = 14.6329$ and $p = 0.000131$, indicating a statistically significant paired difference on this split.

Table 8. McNemar contingency table for run 1

Prediction relationship	Count	Percentage
Both models correct	486	83.79%
Both models wrong	15	2.59%
IndoBERT correct; SVM wrong	57	9.83%
SVM correct; IndoBERT wrong	22	3.79%

Table 9. Representative qualitative errors from run 1

Review excerpt	Rating-derived weak label	Prediction (IndoBERT / SVM)	Assigned aspect(s)	Likely error source
"ongkir 100 juta ... aku hapus aja"	Positive, 4 stars	Negative / Negative	App Quality; Product Quality; Price and Promotion; Delivery Service	The complaint expressed in the review conflicts with the positive rating. The review also activates multiple aspect rules.
"kamera bekas ... tidak bisa dipakai"	Positive, 5 stars	Negative / Negative	App Quality; Delivery Service	Strong negative wording conflicts with the positive rating. The aspect assignment is also potentially ambiguous.
"keranjang belanja tidak bisa dibuka"	Positive, 5 stars	Negative / Negative	App Quality	A clear application failure conflicts with the positive rating-derived label.
"produk banyak tapi stok kosong semua"	Negative, 1 star	Positive / Positive	App Quality; Product Quality	Short informal wording and sparse contextual cues lead both classifiers toward a positive prediction.
"lama sekali donlotnya"	Positive, 5 stars	Negative / Negative	Delivery Service	The term "lama" activates the Delivery Service rule, although the context concerns application downloading. The wording also conflicts with the positive rating.

Note: Weak sentiment labels were derived from star ratings. The aspects were assigned by the rule-based aspect extraction system.

The examples reveal two distinct sources of error: rating-text disagreement and rule-based aspect ambiguity. Reviews with four- or five-star ratings sometimes contain explicit complaints, causing both classifiers to predict negative sentiment against a positive weak label. In addition, polysemous terms may activate an unrelated aspect rule. For example, "lama" in "lama sekali donlotnya" was assigned to Delivery Service even though the context refers to application downloading. Thus, the observed errors reflect noise in both rating-derived sentiment labels and rule-based aspect assignments.

3.6. Discussion

Across the five repeated group hold-out splits in this dataset, the contextual encoder showed higher agreement with rating-derived labels than TF-IDF + SVM. IndoBERT improved mean accuracy by 4.72% and macro F1 by 4.78% relative to TF-IDF + SVM. Its advantage was largest for Delivery Service, where contextual interpretation of temporal and logistical expressions appears more useful than sparse lexical weighting. This pattern is consistent with prior findings that contextual and attention-based models improve aspect-specific sentiment modeling [8], [9], [23].

The significant McNemar result provides paired evidence for run 1, but it should be interpreted together with repeated-run variation rather than as universal superiority. The five splits changed the test groups and produced IndoBERT accuracy from approximately 0.89 to 0.94. Reporting mean, dispersion, and confidence intervals therefore gives a more realistic account than a single high score. The evaluation design also prevents duplicate review text from appearing across training and testing through different aspect rows. The observed consistency applies only to Ruparupa reviews from Google Play and does not establish generalization to other applications or review platforms.

The expanded aspect dictionary substantially broadened the coverage of domain-specific and informal expressions. Nevertheless, rule-based extraction remains vulnerable to keyword polysemy. The word lama may refer to delivery duration or application downloading; gagal may indicate checkout, payment, login, or refund; and rusak may describe a product or packaging. Mixed-aspect reviews and short informal texts can provide sparse or competing cues, while implicit complaints and sarcasm may contain no dictionary term. Accordingly, the aspect counts should be used to describe what the current rule system assigned, rather than to claim verified proportions of user concerns.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Rating-derived polarity creates a second limitation. Several error examples contain strongly negative review text but a 4- or 5-star rating. Conversely, short positive wording may accompany a low rating. A review rating reflects overall satisfaction and does not guarantee that every mentioned aspect shares the same polarity. The classifier may therefore learn rating-associated textual patterns rather than fully human-validated aspect sentiment. Weak supervision reduces annotation cost but introduces label noise and should be supplemented by expert annotation when managerial decisions depend on aspect-level estimates [17].

Excluding three-star reviews created clearer binary weak-label anchors but removed potentially informative neutral, moderate, and mixed evaluations. Such reviews may report usability problems or service details without extreme ratings. The aspect and sentiment distributions therefore describe the retained binary subset rather than the complete prevalence of user concerns in all collected Ruparupa reviews.

Payment Process requires particular caution. It represented only 5.60% of the expanded dataset and produced the highest variability in per-aspect F1. A small aspect group may be more sensitive to split composition and keyword overlap. The strongest defensible practical interpretation is therefore diagnostic: the rule system can help retrieve candidate reviews for closer inspection, while verified prioritization of service improvements requires human-reviewed aspect and sentiment labels. Any managerial use should therefore be limited to exploratory review screening until human-validated aspect and sentiment labels are available.

Future work should construct a manually annotated Indonesian e-commerce ABSA dataset with explicit multi-label aspects and aspect-specific polarity. Two independent annotators, adjudication, and agreement statistics would support reliable evaluation. Supervised multi-label aspect extraction, joint extraction-classification, neutral polarity, implicit aspect handling, and cross-application testing should then be compared with the current transparent rule-based pipeline. Recent Indonesian multi-label IndoBERT research provides a relevant direction for such extensions [24].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

4. CONCLUSION

This study developed a Ruparupa review-analysis pipeline that separates rule-based aspect extraction from IndoBERT-based sentiment classification. The resulting corpus contained 2,873 aspect-review pairs derived from 1,495 unique reviews, and the evaluation used five repeated leakage-free group hold-out splits with separate group-separated validation subsets. IndoBERT achieved mean accuracy of 0.9179, macro F1 of 0.9178, and ROC-AUC of 0.9719, exceeding the calibrated TF-IDF + SVM baseline across all principal metrics. The paired difference was significant on run 1, and IndoBERT produced higher mean F1 for every system-assigned aspect. Within this dataset, IndoBERT showed consistent agreement with rating-derived binary sentiment labels across the five splits. However, the aspect assignments were generated through rule-based matching and were not independently validated by human assessors; ratings also remain weak labels for aspect-specific sentiment. Therefore, the aspect-frequency results and managerial interpretations should be treated as exploratory. Managerial use should be limited to exploratory review screening until human-validated aspect and sentiment labels are available. The findings are limited to one mobile application (Ruparupa), one review source (Google Play), and a binary subset that excludes three-star reviews. Future research should prioritize supervised multi-label aspect extraction and a human-validated Indonesian e-commerce ABSA corpus that includes neutral, mixed, implicit, and sarcastic expressions.

REFERENCES

[1] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, "Analysing app reviews for software engineering: A systematic literature review," *Empir. Softw. Eng.*, vol. 27, art. no. 43, 2022, doi: 10.1007/s10664-021-10065-7.

[2] R. Massenon *et al.*, "Mobile app review analysis for crowdsourcing of software requirements: A mapping study of automated and semi-automated tools," *PeerJ Comput. Sci.*, vol. 10, art. no. e2401, 2024, doi: 10.7717/peerj-cs.2401.

[3] B. Liu, *Sentiment Analysis and Opinion Mining*. in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012. doi: 10.2200/S00416ED1V01Y201204HLT016.

[4] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 11, pp. 11019–11038, 2023, doi: 10.1109/TKDE.2022.3230975.

[5] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland: Association for Computational Linguistics, 2014, pp. 27–35. doi: 10.3115/v1/S14-2004.

[6] M. Pontiki *et al.*, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California: Association for Computational Linguistics, 2016, pp. 19–30. doi: 10.18653/v1/S16-1002.

[7] Y. C. Hua, P. Denny, K. Taskova, and J. Wicker, "A Systematic Review of Aspect-Based Sentiment Analysis: Domains, Methods, and Trends," *Artif. Intell. Rev.*, vol. 57, no. 11, art. no. 296, 2024, doi: 10.1007/s10462-024-10906-z.

[8] D. R. I. M. Setiadi, W. Warto, A. R. Muslikh, K. Nugroho, and A. N. Safriandono, "Aspect-Based Sentiment Analysis on E-commerce Reviews using BiGRU and Bi-Directional Attention Flow," *J. Comput. Theor. Appl.*, vol. 2, no. 4, pp. 470–480, 2025, doi: 10.62411/jcta.12376.

[9] E. Yulianti and N. K. Nissa, "ABSA of Indonesian Customer Reviews Using IndoBERT: Single-Sentence and Sentence-Pair Classification Approaches," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024, doi: 10.11591/eei.v13i5.8032.

[10] S. Imron, E. I. Setiawan, J. Santoso, and M. H. Purnomo, "Aspect Based Sentiment Analysis Marketplace Product Reviews Using BERT, LSTM, and CNN," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 3, pp. 586–591, 2023, doi: 10.29207/resti.v7i3.4751.

[11] M. T. A. Bangsa, S. Priyanta, and Y. Suyanto, "Aspect-Based Sentiment Analysis of Online Marketplace Reviews Using Convolutional Neural Network," *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 14, no. 2, pp. 123–134, 2020, doi: 10.22146/ijccs.51646.

[12] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[14] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China, 2020, pp. 843–857.

[15] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[16] E. Mulyati, M. I. C. Rachmatullah, and A. S. Firmansyah, "Sentiment Analysis of Pospay Application Reviews Using the BERT Deep Learning Method," *J. Tek. Inform.*, vol. 18, no. 2, pp. 173–183, 2025, doi: 10.15408/jti.v18i2.41116.

[17] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid Training Data Creation with Weak Supervision," *VLDB J.*, vol. 29, pp. 709–730, 2020, doi: 10.1007/s00778-019-00552-1.

[18] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science," *Patterns*, vol. 4, no. 9, art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[19] S. Kapoor *et al.*, "REFORMS: Consensus-Based Recommendations for Machine-Learning-Based Science," *Sci. Adv.*, vol. 10, no. 18, art. no. eadk3452, 2024, doi: 10.1126/sciadv.adk3452.

[20] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.

[21] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[22] F. Pedregosa *et al.*, "Scikit-Learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[23] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-Based Sentiment Analysis Using BERT," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland: Linköping University Electronic Press, 2019, pp. 187–196.

[24] H. N. Alfiana, A. Doewes, and B. Widoyono, "Aspect-Based Sentiment Analysis of Access by KAI Application Reviews Using IndoBERT for Multi-Label Classification Tasks," *J. Tek. Inform.*, vol. 7, no. 1, pp. 286–306, 2026, doi: 10.52436/1.jutif.2026.7.1.5402.