Intechno Journal :
Information Technology Journal
Vol. 8 No. 1 July 2026
Homepage: https://jurnal.amikom.ac.id/index.php/intechno
INTECHNO Journal
INFORMATION TECHNOLOGY JOURNAL
e-ISSN: 2655-1438
p-ISSN: 2655-1632

Aspect-Level User Sentiment Patterns in Shopee Mobile App Reviews Using TF-IDF and Logistic Regression

Itishom Al Khoiry
Department of Digital Business, Universitas Persatuan Guru Republik Indonesia Semarang, Indonesia

Article Info

Article history:

Received: 29 June 2026
Revised: 13 July 2026
Accepted: 14 July 2026
Published: 31 July 2026

Keywords:

Aspect-level analysis
Logistic regression
Mobile app reviews
Sentiment classification
TF-IDF

ABSTRACT

Purpose: This study examines recent Shopee mobile application reviews from a service diagnostic perspective by linking sentiment classification with aspect-level analysis. The aim is to identify not only whether user feedback is positive or negative, but also which service dimensions are associated with satisfaction and dissatisfaction.

Methods: A total of 160,540 Google Play Store reviews of the Shopee Android application were collected from 1 January to 27 June 2026. After cleaning, rating-based sentiment labeling, and preprocessing, all 151,392 eligible reviews were used for modeling. TF-IDF unigram and bigram features were fitted inside a pipeline and combined with Logistic Regression as the proposed model, alongside a majority class baseline, Naive Bayes, and Linear SVM. Evaluation used accuracy, Macro F1-score, 5-fold stratified cross-validation, and the public PRDECT-ID benchmark. A keyword-based aspect mapping then interpreted sentiment across six service dimensions.

Result: TF-IDF with Logistic Regression achieved 0.9247 accuracy and 0.9155 Macro F1-score on the test set, and 0.9175 Macro F1-score under cross-validation. On PRDECT-ID it reached 0.9470 Macro F1-score. Positive sentiment concentrated in Promotion and Voucher, Product and Seller, and Customer Service, while negative sentiment concentrated in Delivery, Application Performance, and Payment. Advertising and delivery-related expressions were the strongest negative signals.

Novelty: This study proposes a service diagnostic task with aspect-level analysis, which includes sentiment classification, term-weight interpretation, and service-aspect mapping, to position Shopee review analysis, and validates the interpretable pipeline on an external public benchmark. The strategy transforms bulk app reviews into actionable data to track platform service quality.

Corresponding Author:

Itishom Al Khoiry
Email: itishom@upgris.ac.id

1. INTRODUCTION

Mobile commerce has become a primary channel of digital consumption in Indonesia and across Southeast Asia, and shopping applications now mediate a large share of everyday retail activity. The quality of the mobile app itself, such as performance, payment flow, delivery handling, and customer service, is now playing a key role in keeping users. Shopee is among the most popular marketplace apps in the region, and the reviews users leave on the Google Play Store are a constant flow of unprompted comments on these aspects of the service [1], [2].

App store reviews are a valuable source of user-generated content, as they are a numerical rating along with free-text commentary, which is written voluntarily by real users in real usage contexts. Previous research in requirements engineering has demonstrated that such feedback can be used to facilitate communication between users and developers, but a significant amount of it is noise and needs to be filtered before it can be informative [3]–[5]. The number of incoming reviews is too large to be processed manually by a platform of Shopee's size, and automated text classification is a solution to transform raw reviews into structured satisfaction signals.

Sentiment classification is the standard technique for this task. It labels each review with a polarity to monitor the level of positive and negative opinions over time [6]. A common and reproducible representation for review text is Term Frequency Inverse Document Frequency (TF-IDF), which weights terms by their local frequency and global rarity, paired with linear classifiers such as Naive Bayes, Logistic Regression, and Support Vector Machines [7], [8]. These models remain attractive for operational monitoring because they are fast to train, transparent, and easy to interpret through their term weights.

In addition to a polarity score, practitioners are increasingly required to be aware of which service area a complaint/compliment is related to. To tackle this, aspect-level analysis relates sentiment to individual aspects of the service, such as delivery, payment, or product quality [9], [10]. Fine-grained sentiment analysis of app reviews has been proven to uncover actionable, feature-specific sentiment that is lost in a global sentiment score [9], and aspect-based sentiment analysis has been used in e-commerce app reviews to uncover which features lead to satisfaction and dissatisfaction [11], [12]. However, full supervised aspect-based sentiment analysis typically requires manually annotated aspect and polarity labels, which are costly to produce at scale.

Previous studies have analyzed app store and e-commerce reviews using machine learning and deep learning models, including Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, ensemble methods, and neural approaches [13]–[19]. A systematic review of app review analysis for software engineering shows that most published pipelines extract user opinions, feature requests, and defect reports, and that classical supervised models remain widely used because they are inexpensive to retrain and easy to audit [3]. Dąbrowski et al. [4] extend this evidence through evaluation and replication studies of app review mining tools and report that classification performance is highly sensitive to the dataset and to the labeling procedure adopted, which is why the class distribution, the provenance of the labels, and imbalance-aware metrics should be reported rather than accuracy alone. At a larger scale, Qureshi et al. [5] evaluate several machine learning models on 251,661 Google Play reviews represented with TF-IDF and find that classical classifiers remain strong baselines for app review sentiment. Feature-level studies further show that user dissatisfaction concentrates in a small number of app functions rather than being distributed uniformly [9]. Most Shopee-specific studies, however, remain focused on sentiment polarity and overall accuracy [1], [7], while aspect-based sentiment studies suggest that user opinion becomes more actionable when it is linked to specific service attributes rather than reported only as a general positive or negative label [8]–[12].

However, sentiment analysis of marketplace reviews is often limited to reporting overall polarity, which restricts its usefulness for service management. Accuracy-based evaluation can also be insufficient when the data are imbalanced, because a high accuracy may conceal weak performance on the minority class. In addition, a model evaluated only on a self-collected corpus cannot be compared directly with earlier work, so its generalizability remains unverified. This study therefore proposes an interpretable service monitoring pipeline for Shopee review analysis. It adopts TF-IDF with Logistic Regression as the proposed approach, trains a majority class baseline, Multinomial Naive Bayes, and Linear SVM on identical features as comparative classifiers, reports Macro F1-score and stratified cross-validation instead of accuracy alone, validates the approach on the public PRDECT-ID benchmark of Indonesian e-commerce reviews [20], explains the selected model through its term weights, and adds a keyword-based aspect-level analysis that identifies the service areas receiving positive and negative feedback.

The purposes of this study are as follows: (1) to classify large-scale Shopee mobile app reviews using TF-IDF features with Logistic Regression as the proposed interpretable model, with a majority class baseline, Naive Bayes, and Linear SVM as comparative classifiers; (2) to evaluate the models with metrics that account for class imbalance, namely Macro F1-score and stratified cross-validation; (3) to verify the generalizability of the approach on a public Indonesian e-commerce review benchmark; and (4) to examine the distribution of positive and negative sentiment across specific service aspects, namely application performance, delivery, payment, product and seller, customer service, and promotion and voucher. The contribution lies in moving beyond general polarity classification to link user sentiment with specific service areas, with the aspect-level component serving as a descriptive diagnostic layer rather than a supervised classifier.

2. METHOD

This study adopts a quantitative text mining pipeline consisting of eight steps: Data Collection, Data Cleaning and Sentiment Labeling, Text Pre-processing, TF-IDF Feature Extraction, Data Splitting, Model Training and Evaluation, External Benchmark Evaluation, and a supplementary K31-42eyword-based Aspect-Level Analysis. The entire pipeline is implemented in Python with the scikit-learn library, and the description below is restricted to the operations performed in the research notebook so that the work can be reproduced.

2.1 Data Collection

Reviews were collected from the Google Play Store listing of the Shopee Android application (application identifier: com.shopee.id) using an automated review scraper. The dataset covers reviews published from 1 January 2026 to 27 June 2026. After applying the date filter, 160,540 reviews were retained for analysis.

2.2 Data Cleaning and Sentiment Labeling

The rating (score), review text (content), and timestamp (at) fields were kept from the raw reviews, along with optional metadata fields if they were available. Records that did not contain a rating, text, or timestamp were excluded, duplicate records of text, rating, and timestamp were removed, empty reviews were deleted, and the data were limited to the study period, resulting in 160,538 reviews. Sentiment labels were derived directly from the user rating: ratings of 1 and 2 were labeled negative and ratings of 4 and 5 were labeled positive. To avoid ambiguity in a rating-based labeling scheme, reviews rated 3 were excluded from the classification task, resulting in 5,539 reviews being removed and 154,999 labeled reviews and a binary labeling scheme. The labels generated are thus proxy labels, not ground truth labels, because they are based on ratings. The cost of rating-based labelling is low, it is fully reproducible and can be applied at this scale, but the text of a review can disagree with the rating that is attached to it, and any disagreement is treated as label noise within the model, which is discussed and reflected in the external benchmark evaluation using human-annotated labels. The labeling scheme also results in an unbalanced binary distribution of about 2 positive reviews per negative review, which is reported in the results and is the reason why the model is selected and reported using Macro F1-score, instead of accuracy.

2.3 Text Preprocessing

Each review was processed through a standardized text preprocessing procedure. The text was converted to lowercase, and URLs, user mentions, hashtags, non-alphabetic characters, and repeated whitespace were removed. The cleaned text was then tokenized and normalized using an Indonesian slang dictionary, for example converting gk, ga, and gak to tidak, and apk or app to aplikasi. Indonesian stopwords were removed, but negation terms such as tidak, tak, bukan, belum, jangan, ga, gak, nggak, and enggak were retained because they are important sentiment cues. Single-character tokens were also discarded. Sastrawi stemming was not applied in the final experiment in order to preserve discriminative word forms commonly found in user reviews and to reduce preprocessing time [21].

2.4 TF-IDF Feature Extraction

The preprocessed text was converted into numerical features using TF-IDF weighting. The vectorizer used unigram and bigram features, a maximum vocabulary of 30,000 features, a minimum document frequency of 2, and sublinear term-frequency scaling. TF-IDF emphasizes terms that are frequent within a review but relatively rare across the corpus [7], [8]. Equation (1) gives the TF-IDF weight of term (t) in document (d), where $(tf_{t,d})$ is the frequency of term (t) in document (d), (N) is the total number of documents, and (df_t) is the number of documents containing term (t).

$$w_{t,d} = \left(1 + \log\left(tf_{t,d}\right)\right) \times \left[\log\left(\frac{1+N}{1+df_t}\right) + 1\right] \quad (1)$$

2.5 Data Splitting

The complete set of reviews remaining after preprocessing, comprising 151,392 reviews, was used for modeling, so the reported performance refers to the entire eligible corpus. The dataset was split into training and test sets using an 80:20 stratified split with a fixed random seed of 42, which preserves the class proportions in both partitions and makes the split reproducible. The split produced 121,113 training reviews and 30,279 test reviews.

2.6 Model Training and Evaluation

The proposed approach of this study is TF-IDF with Logistic Regression, which was pre-determined as the main model since the coefficients of Logistic Regression provide an easy and transparent interpretation of the terms of each sentiment class. Three additional models were trained on the same TF-IDF features, so that any performance difference is due to the classifier and not the representation: a majority class baseline which always predicts the most common label; Multinomial Naive Bayes with a smoothing parameter of 0.5; and Linear SVM with C set to 1.0. Logistic Regression was run with a maximum of 2,000 iterations and a regularization parameter C of 1.0. The three models are not competing models, but comparative baselines: the

majority class baseline is the minimum performance level that any useful model should not fall below, and Naive Bayes and Linear SVM are statements about the possibility of an alternative linear model being materially better on the same features. The held-out test set was used to evaluate each model by accuracy, macro-averaged and weighted-averaged precision, recall and F1-score, as well as per-class precision, recall and F1-score. The macro F1 score is used as the main measure as it gives equal weight to both classes and is suitable for an imbalanced corpus [11]. The primary model was also evaluated using 5-fold stratified cross validation on the entire dataset to ensure the stability of the model performance and a confusion matrix was generated to understand the error pattern.

2.7 External Benchmark Evaluation

To allow comparison with previous work under conditions that are not specific to the self-collected corpus, the same pipeline was evaluated on PRDECT-ID, a public dataset of 5,400 Indonesian product reviews collected from 29 Tokopedia product categories and released with binary positive and negative sentiment labels [20]. The released distribution is 2,821 negative and 2,579 positive reviews, so the benchmark is close to balanced and its labels are annotated by human raters rather than derived from ratings, which makes it a useful external check on the labeling scheme used for the Shopee corpus. Exact duplicate review texts were removed before splitting so that identical texts could not appear in both partitions: 95 duplicates were removed and 2 reviews became empty after preprocessing, leaving 5,303 unique reviews (2,751 negative and 2,552 positive). Because the dataset is distributed without official partitions, the same stratified 80:20 split with a random seed of 42 was applied, producing 4,242 training and 1,061 test reviews, and the same four models were trained on identical TF-IDF features. The held-out evaluation was supplemented with 5-fold stratified cross-validation of the primary model on all 5,303 reviews. The text preprocessing function applied to the benchmark was identical to the one applied to the Shopee corpus.

2.8 Supplementary Keyword-Based Aspect-Level Analysis

To interpret the substance of the feedback, a supplementary aspect-level analysis assigned each review to one of six service aspects using predefined Indonesian keyword lists: Application Performance, Delivery, Payment, Product and Seller, Customer Service, and Promotion and Voucher. For each review, the number of keyword matches per aspect was counted, and the review was assigned to the aspect with the most matches; reviews with no keyword match were assigned to an Other category. This mapping is a descriptive, rule-based grouping. It is not a trained aspect-based sentiment model and does not assign sentiment per aspect independently; aspect-level sentiment is obtained by cross-tabulating the rule-assigned aspect with the rating-based sentiment label.

3. RESULTS AND DISCUSSION

3.1 Dataset Preparation

Table 1 summarizes the number of reviews retained at each preparation stage, and Figure 1 presents the same information as a data preparation flow. Scraping returned 160,540 reviews within the research period. Removing records with missing fields, exact duplicates, and empty texts, and restricting the data to the study period, left 160,538 reviews. Excluding the 5,539 reviews rated 3 left 154,999 labeled reviews, and after preprocessing and the removal of reviews that became empty, 151,392 reviews remained. All 151,392 reviews constitute the modeling dataset. Figure 1 makes the effect of each filter explicit: cleaning removes almost nothing, which indicates that the scraped corpus contained few malformed or duplicated records, while the two substantive reductions come from the exclusion of rating-3 reviews (3.5 percent of the cleaned corpus) and from preprocessing (a further 2.3 percent). The figure therefore documents that 94.3 percent of the scraped reviews survive into the analysis, so the modeling dataset is not a narrow or selective subset of the collected data. The monthly distribution of positive and negative reviews across the period is shown in Figure 2.

Table 1. Number of reviews retained at each data preparation stage

Stage	Number of reviews
Initial scraped reviews	160,540
After cleaning (missing values, duplicates, empty text, date filter)	160,538
After rating-3 exclusion	154,999
After text preprocessing (final modeling dataset)	151,392

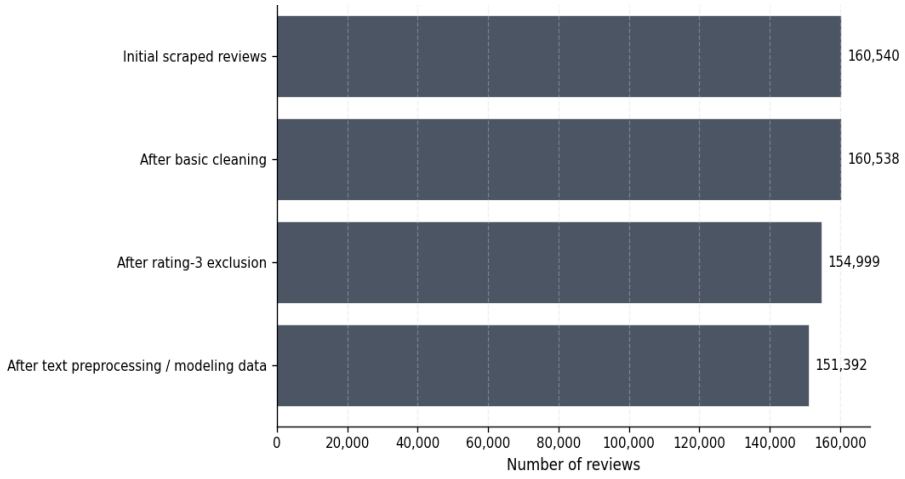

Figure 1. Data preparation flow

3.2 Sentiment Distribution

The binary sentiment distribution of the modeling dataset is presented in Table 2. The positive class accounts for 100,809 reviews (66.6 percent) and the negative class for 50,583 reviews (33.4 percent), an imbalance of approximately two to one that confirms the use of Macro F1-score as the primary selection metric. The imbalance is material: a classifier that always predicted the positive class would already obtain 66.6 percent accuracy while contributing nothing on the minority class, as the majority class baseline in Table 3 confirms. Figure 2 shows the monthly share of positive and negative reviews. Positive reviews dominated in every month from January to May 2026, when negative reviews accounted for between 22 and 28 percent of the monthly volume. In June 2026 the pattern reversed: 22,033 negative reviews were recorded against 13,445 positive reviews, so negative reviews represented 62.1 percent of that month. This reversal is concentrated in a single month and is examined further in the term-weight and aspect-level analyses.

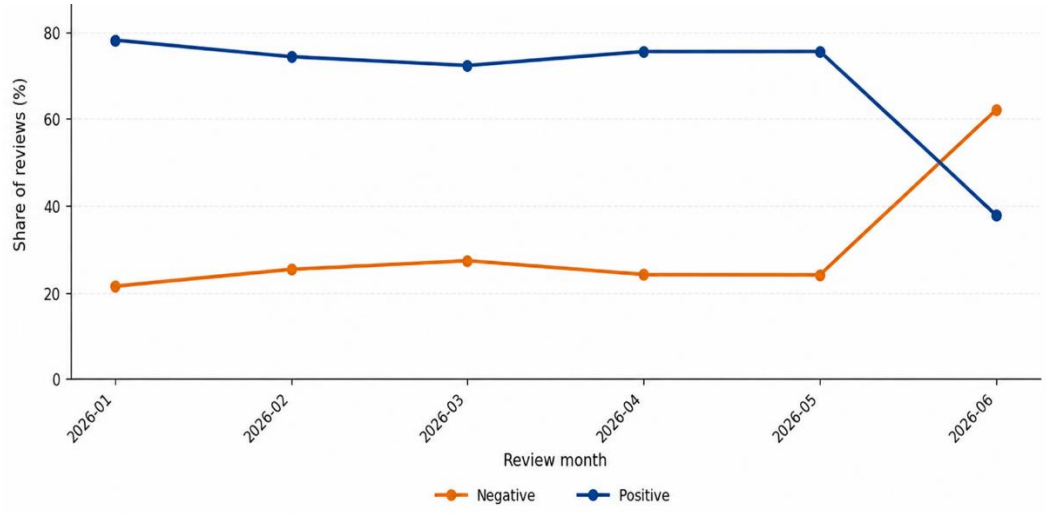

Figure 2. Monthly trend of positive and negative Shopee reviews

Table 2. Binary sentiment distribution in the modeling dataset

Sentiment	Count	Proportion
Positive	100,809	0.666
Negative	50,583	0.334
Total	151,392	1.000

3.3 Model Performance

Table 3 reports the test-set performance of the four models on the 30,279-review test set. The majority class baseline records the lowest Macro F1-score (0.3997) because it predicts only the majority positive class and ignores the minority class entirely, although its accuracy of 0.6659 illustrates how misleading accuracy alone would be on this corpus. All three trained classifiers improve substantially over the baseline, which confirms that the TF-IDF features carry a strong polarity signal. The differences among the trained models are

small: Naive Bayes and Linear SVM reach Macro F1-scores of 0.9106 and 0.9100 respectively, while the proposed TF-IDF and Logistic Regression model attains both the highest accuracy (0.9247) and the highest Macro F1-score (0.9155). Logistic Regression is therefore retained as the primary model, and the comparison shows that the proposed classifier is not outperformed by the alternative linear models on identical features. Figure 3 compares the models in terms of accuracy, Macro F1, Macro precision, and Macro recall.

Table 3. Test-set performance of the evaluated models

Model	Accuracy	Prec. (macro)	Rec. (macro)	F1 (macro)	F1 (weighted)
Majority Class Baseline	0.6659	0.3329	0.5000	0.3997	0.5323
Naive Bayes	0.9193	0.9053	0.9168	0.9106	0.9198
Logistic Regression	0.9247	0.9151	0.9159	0.9155	0.9248
Linear SVM	0.9199	0.9103	0.9097	0.9100	0.9199

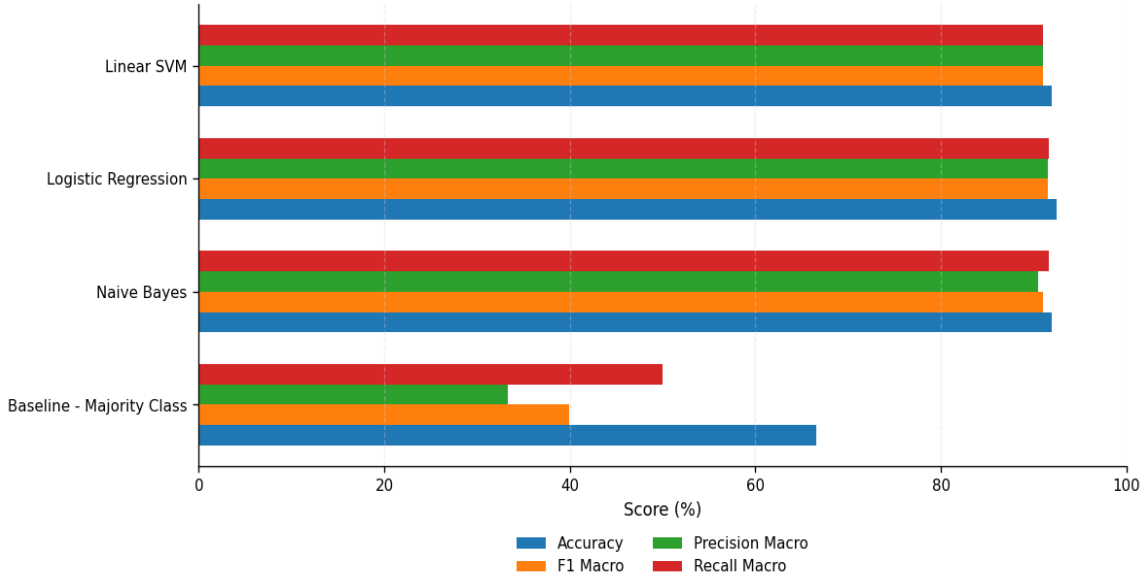

Figure 3. Model performance summary

Logistic Regression is therefore selected for detailed reporting. The selection rests on three grounds. It records the highest Macro F1-score (0.9155) and the highest accuracy (0.9247) among the compared classifiers, so no accuracy is sacrificed by preferring it. Its margin over Naive Bayes and Linear SVM is small (0.0049 and 0.0055 Macro F1 respectively), which means the three classifiers extract comparable information from the same TF-IDF features, and in that situation the deciding criterion is interpretability: unlike Naive Bayes, whose log-probabilities reflect class frequency as well as term discriminativeness, the coefficients of Logistic Regression can be read directly as the weight of each term for each sentiment class, which is what the aspect-level diagnostic objective of this study requires. Finally, Logistic Regression provides probability estimates for continuous sentiment monitoring, unlike Linear SVM. The classification report of the primary model, Logistic Regression, is given in Table 4 and its confusion matrix in Table 5. The model separates the two classes well: the positive class reaches an F1-score of 0.943 and the negative class an F1-score of 0.888, with the lower negative score reflecting the smaller support for that class (10,117 reviews against 20,162). Figure 4 presents the confusion matrix graphically.

Table 4. Classification report of the best model (Logistic Regression)

Class	Precision	Recall	F1-score	Support
Negative	0.886	0.889	0.888	10,117
Positive	0.944	0.943	0.943	20,162
Macro avg	0.915	0.916	0.916	30,279
Weighted avg	0.925	0.925	0.925	30,279

Table 5. Confusion matrix of the best model (Logistic Regression)

	Predicted Negative	Predicted Positive
Actual Negative	8,997	1,120
Actual Positive	1,159	19,003

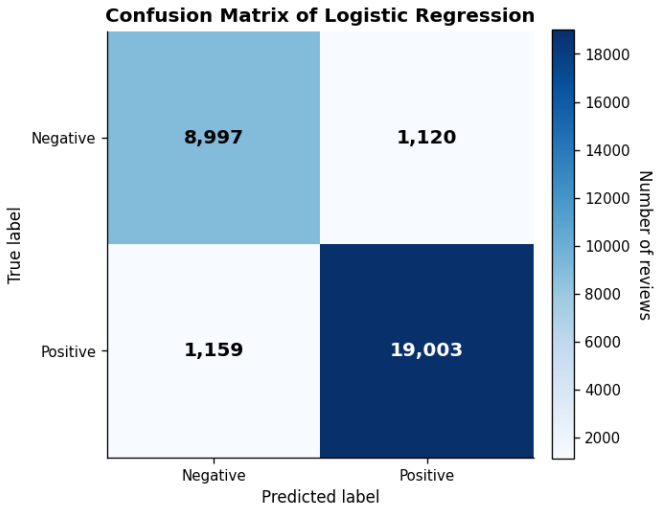

Figure 4. Confusion matrix of the best model

The confusion matrix shows that the model produced 1,120 false positives and 1,159 false negatives out of 30,279 test reviews. Although the two error types are similar in absolute number, the class-normalized error rate is higher for Negative reviews, consistent with their smaller representation in the training data.

3.4 Cross Validation

To confirm that the performance of Logistic Regression is stable and not an artifact of a single split, 5-fold stratified cross-validation was performed on the complete dataset of 151,392 reviews. Table 6 reports the mean and standard deviation of each metric across folds. All standard deviations are below 0.002, indicating that the model generalizes consistently across partitions, which matters here because the dataset is imbalanced and a reliable model must hold its performance on both classes. The cross-validated Macro F1-score of 0.9175 is close to the single-split test value of 0.9155, so the reported result does not depend on one particular partition of the data.

Table 6. Five-fold cross-validation results for the best model

Metric	Mean	Std
Accuracy	0.9264	0.0007
Precision (macro)	0.9164	0.0008
Recall (macro)	0.9187	0.0012
F1 (macro)	0.9175	0.0008
F1 (weighted)	0.9265	0.0007

3.5 Discriminative Terms

To make the classifier interpretable, the coefficients of the Logistic Regression model were examined. Figure 5 displays the strongest positive and negative terms. The most negative coefficients are dominated by advertising-related and evaluative words: *iklan* (advertisement, -7.81) carries the strongest negative weight, followed by *jelek* (bad, -7.76), *tidak* (not, -7.56), *buruk* (poor, -7.18), *iklannya* (its advertisement, -6.19), and *sampah* (rubbish, -4.85). Complaint-specific terms also appear, including *maksa* (forced), *boikot* (boycott), the courier name *spx* (-4.12), and the bigram *tidak sesuai* (not as described). The strongest positive coefficients are *membantu* (helpful, 8.49), *mudah* (easy, 7.15), *cepat* (fast, 6.32), *puas* (satisfied, 6.07), *terbaik* (best, 5.34), *belanja* (shopping, 4.81), and *murah* (cheap, 4.73). The bigram features are informative in their own right: while the unigram *tidak* is a strong negative cue, the bigrams *tidak mengecewakan* (not disappointing, 6.60), *tidak ribet* (not complicated, 5.99), and *tidak kecewa* (not disappointed, 5.32) carry large positive weights, which shows that the unigram and bigram representation captures negated praise that a unigram-only model would misread. These weights give a transparent account of why a review is classified as it is, and they indicate

that in-app advertising and delivery handling were the leading lexical drivers of negative sentiment during the period, while ease of use, speed, and value drove positive sentiment.

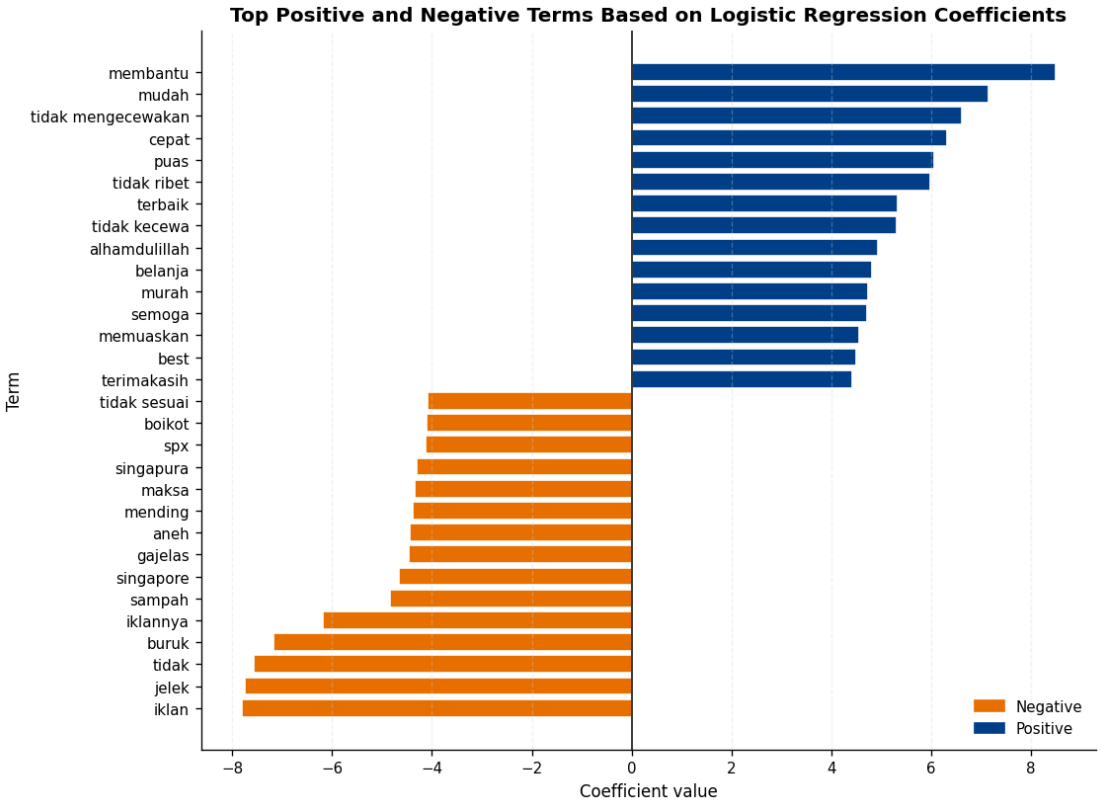

Figure 5. Top positive and negative terms based on Logistic Regression coefficients

3.6 Aspect-Level Sentiment Findings

The 151,392 reviews were distributed across six service aspects, with the remaining 79,864 reviews that did not include any aspect keyword being assigned to Other. Table 7 cross tabulates aspect by sentiment, and Figure 6 displays the positive and negative share for each substantive aspect. The Other category is large, with 79,864 reviews (52.75 percent of the corpus), which is due to the brevity and generality of many app store reviews, and is omitted from Figure 6 so that the figure focuses on the interpretable aspects. Among the substantive aspects, Promotion and Voucher is the most positive (75.85 percent positive), followed by Product and Seller (64.57 percent positive) and Customer Service (62.66 percent positive). The areas of Delivery (55.45 percent negative), Application Performance (54.46 percent negative), and Payment (53.59 percent negative) are mostly negative. Application Performance is also the highest by volume (28,977 reviews), followed by Delivery (17,435 reviews), meaning that app stability and delivery handling are the most discussed and most negatively perceived service areas.

Table 7. Aspect-level sentiment distribution (keyword-based mapping)

Aspect	Negative	Positive	Total
Application Performance	15,781	13,196	28,977
Delivery	9,668	7,767	17,435
Product and Seller	4,532	8,258	12,790
Customer Service	1,837	3,082	4,919
Promotion and Voucher	935	2,936	3,871
Payment	1,895	1,641	3,536
Other	15,935	63,929	79,864
Total	50,583	100,809	151,392

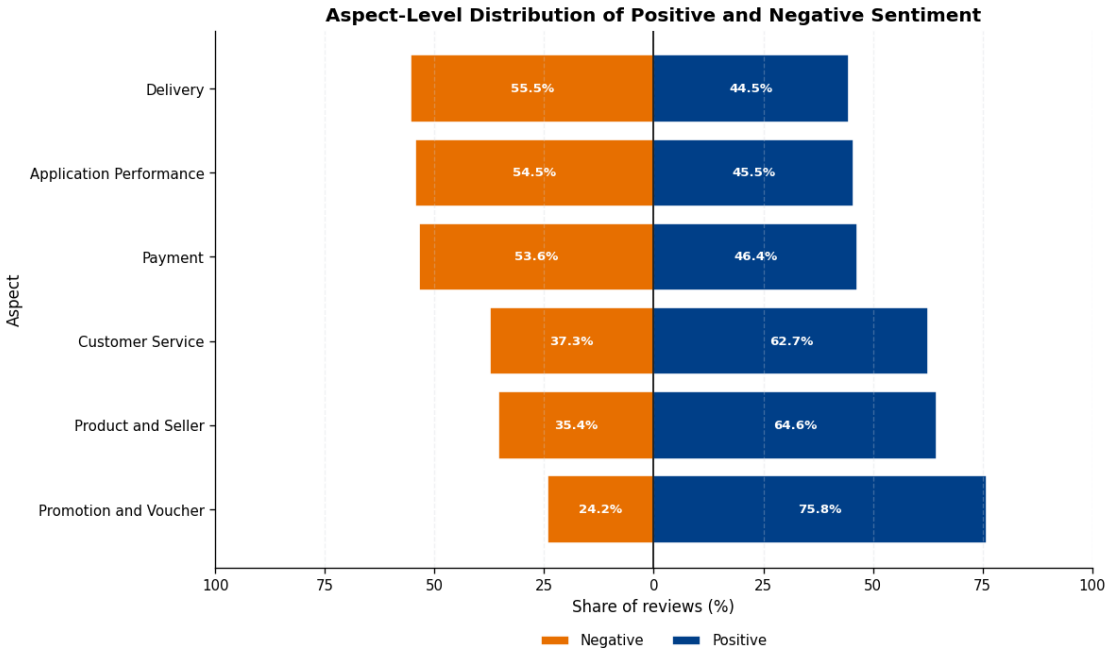

Figure 6. Aspect-level distribution of positive and negative sentiment

3.7 External Benchmark Evaluation on PRDECT-ID

To assess generalizability beyond the self-collected corpus, the same pipeline was evaluated on the PRDECT-ID benchmark. Table 8 presents the test results on 1,061 reviews. TF-IDF with Logistic Regression achieved 0.9472 accuracy and 0.9470 Macro F1-score, outperforming Naive Bayes (0.9422 Macro F1) and performing comparably to Linear SVM (0.9498 Macro F1). The 0.28 percentage point Macro F1 difference is smaller than the fold-to-fold variation, indicating no meaningful performance difference. The majority class baseline achieved only 0.3414 Macro F1, as expected for a near-balanced dataset.

Table 8. Held-out test performance on the PRDECT-ID benchmark

Model	Accuracy	Prec. (macro)	Rec. (macro)	F1 (macro)	F1 (weighted)
Majority Class Baseline	0.5184	0.2592	0.5000	0.3414	0.3540
Naive Bayes	0.9425	0.9450	0.9413	0.9422	0.9424
Logistic Regression	0.9472	0.9497	0.9460	0.9470	0.9471
Linear SVM	0.9500	0.9521	0.9490	0.9498	0.9500

Table 9 reports the 5-fold stratified cross-validation of Logistic Regression on all 5,303 benchmark reviews. The mean Macro F1-score is 0.9405 with a standard deviation of 0.0090, which confirms the stability of the held-out result.

Table 9. Five-fold cross-validation of Logistic Regression on the PRDECT-ID benchmark

Metric	Mean	Std
Accuracy	0.9408	0.0089
Precision (macro)	0.9428	0.0085
Recall (macro)	0.9397	0.0091
F1 (macro)	0.9405	0.0090
F1 (weighted)	0.9407	0.0089

The benchmark score is higher than that obtained on the Shopee corpus (0.9470 versus 0.9155 Macro F1-score). PRDECT-ID provides a complementary evaluation setting because it is nearly balanced, covers multiple product categories, and uses sentiment labels corresponding to expert-guided emotion annotations. In contrast, the Shopee corpus consists of short app reviews with rating-derived proxy labels and a stronger class imbalance. Despite these differences, Logistic Regression maintained strong performance across both datasets. This consistency supports the robustness and generalizability of the pipeline beyond the self-collected corpus while providing an objective external reference under different dataset conditions.

3.8 Discussion and Practical Implications

The findings have direct implications for service monitoring on a marketplace platform, and they also position the proposed approach relative to previous work. The combination of TF-IDF with a linear classifier converts a continuous stream of Shopee reviews into a tracked polarity signal at low computational cost and with fully auditable term weights. The Macro F1-score of 0.9155 obtained here is consistent with the performance reported for classical models on large-scale Google Play review corpora [5] and for Shopee and e-commerce reviews classified with TF-IDF and Naive Bayes [1], [2], [7], and the external benchmark result of 0.9470 Macro F1-score on PRDECT-ID confirms that the pipeline is competitive on a public Indonesian e-commerce dataset. The present study nevertheless departs from that literature in three respects. First, earlier Shopee studies report accuracy on imbalanced corpora, whereas this study selects and reports the model on Macro F1-score and confirms it with cross-validation on the full corpus, so performance on the minority negative class is not concealed by the majority class. Second, studies that report higher headline scores generally rely on deep learning or ensemble architectures [13]–[16]; the present results do not contradict them, and the contribution here is that a transparent linear model reaches a comparable level on the same task while remaining interpretable at the level of individual terms, which is the property that makes the output usable for service diagnostics. Third, the aspect-level layer is descriptive rather than supervised, in contrast with the annotated aspect-based sentiment pipelines of [8], [11], [12]; it therefore trades per-aspect validity for the ability to operate at scale without manual annotation.

The aspect-level analysis provides the operational value of the study. The analysis breaks down feedback into Application Performance, Delivery, Payment, Product and Seller, Customer Service, and Promotion and Voucher, highlighting the areas of negative sentiment and thus the areas where remediation is most likely to impact overall satisfaction. On balance, Promotion and Voucher, Product and Seller, and Customer Service are positive, whereas Delivery, Application Performance, and Payment are negative. The term-level evidence points in the same direction: the terms that are related to advertising (iklan, iklannya) have the highest negative weights, followed by the terms related to delivery (spx, courier, etc.), and the terms that indicate ease, speed, and value (membantu, mudah, cepat, puas, murah) have the highest positive weights. The convergence between the coefficient view and the keyword view is what makes the reversal observed in June 2026 interpretable: the two aspects that dominate the corpus by volume, application performance and delivery, are also among the aspects with the highest negative shares. The finding that dissatisfaction concentrates in a small number of functions rather than spreading evenly across the application is consistent with feature-level studies of app reviews [9], [12], [14], [15]. As in that literature, the aspect view is descriptive and is intended to direct attention rather than to serve as a validated per-aspect classifier.

Three cautions apply to the interpretation. First, the associations reported here are correlational, and no aspect can be said to cause a change in satisfaction. Second, the sentiment labels are derived from star ratings rather than from manual annotation, so reviews whose text disagrees with the accompanying rating enter the training data as label noise; the higher score obtained on the human-annotated PRDECT-ID benchmark is consistent with this explanation. Third, the classifier was trained and evaluated on Shopee Google Play reviews from a single six-month window, and its performance should not be assumed to transfer to other platforms, app stores, or periods without retraining. Within these limits, the pipeline provides an actionable and repeatable instrument for continuous review monitoring.

4. CONCLUSION

This study classified Shopee mobile application reviews from the Google Play Store using TF-IDF features with Logistic Regression and interpreted the results at the level of service aspects. Three findings emerge. First, the proposed model classifies review polarity reliably on an imbalanced corpus, reaching a Macro F1-score of 0.9155 on the held-out test set and 0.9175 under 5-fold cross-validation on all 151,392 eligible reviews, and it is not outperformed by Naive Bayes or Linear SVM trained on identical features. Second, the approach transfers to an independent, human-annotated Indonesian e-commerce benchmark, reaching a Macro F1-score of 0.9470 on PRDECT-ID, supporting the generalizability of the approach beyond the self-collected corpus. Third, negative sentiment in the corpus is not diffuse: it concentrates in Delivery, Application Performance, and Payment, while Promotion and Voucher, Product and Seller, and Customer Service are positive on balance, and the coefficient analysis identifies in-app advertising and delivery handling as the strongest lexical drivers of dissatisfaction.

The practical implication is that an interpretable linear classifier is sufficient to convert a continuous stream of app store reviews into a service diagnostic signal. Because the model exposes its decision rule as term

weights, platform teams can trace a movement in the sentiment series back to the specific expressions that produced it, and the aspect-level analysis indicates which service function should absorb the remediation effort. In this corpus, the reversal to a negative majority in June 2026, together with the large negative coefficients on advertising terms, points to in-app advertising and delivery handling as the areas where intervention would most plausibly affect overall satisfaction.

The findings should be interpreted within the scope of the study design. The sentiment labels were generated from the star ratings, which gave a consistent and scalable labelling method for the whole review corpus, but some differences in the opinion expressed in the text and the rating may exist. The aspect layer based on the keyword was used to provide a transparent interpretation of the explicitly mentioned service dimensions, while 52.75% of the reviews were classified as Other, as they included general feedback or did not correspond to the six predefined aspects. Furthermore, the analysis is based on one application, one app store, and a six-month timeframe, so the sentiment change in June 2026 is a significant pattern that needs to be studied further. This framework can be further strengthened by using a subset of manually labelled aspect and polarity, comparing the interpretable pipeline with other Indonesian language models that are trained on the context, and expanding the analysis to other platforms and time periods.

DECLARATION OF COMPETING INTERESTS

The authors have no conflicts of interest to declare.

DATA AVAILABILITY

The data used in this study were collected by scraping publicly available user reviews from the Google Play Store and are available from the corresponding author upon reasonable request.

REFERENCES

[1] D. Pratmanto, R. Rousyati, F. F. Wati, A. E. Widodo, S. Suleman, and R. Wijianto, "App Review Sentiment Analysis Shopee Application in Google Play Store Using Naive Bayes Algorithm," in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 12043.

[2] M. J. Hossain, D. Das Joy, S. Das, and R. Mustafa, "Sentiment analysis on reviews of e-commerce sites using machine learning algorithms," in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, IEEE, 2022, pp. 522–527.

[3] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, "Analysing app reviews for software engineering: a systematic literature review," *Empir. Softw. Eng.*, vol. 27, no. 2, p. 43, 2022.

[4] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, "Mining and searching app reviews for requirements engineering: Evaluation and replication studies," *Inf. Syst.*, vol. 114, p. 102181, 2023.

[5] A. A. Qureshi, M. Ahmad, S. Ullah, M. N. Yasir, F. Rustam, and I. Ashraf, "Performance evaluation of machine learning models on large dataset of android applications reviews," *Multimed. Tools Appl.*, vol. 82, no. 24, pp. 37197–37219, 2023.

[6] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, 2022.

[7] R. Kosasih and A. Alberto, "Sentiment analysis of game product on shopee using the TF-IDF method and naive bayes classifier," *Ilk. J. Ilm.*, vol. 13, no. 2, pp. 101–109, 2021.

[8] N. P. Arthamevia and M. D. Purbolaksono, "Aspect-based sentiment analysis in beauty product reviews using tf-idf and svm algorithm," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2021, pp. 197–201.

[9] Y. Wang, J. Wang, H. Zhang, X. Ming, L. Shi, and Q. Wang, "Where is your app frustrating users?," in *Proceedings of the 44th international conference on software engineering*, 2022, pp. 2427–2439.

[10] 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, 2022.

[11] P. Hajek, L. Hikkerova, and J.-M. Sahut, "Fake review detection in e-Commerce platforms using aspect-based sentiment analysis," *J. Bus. Res.*, vol. 167, p. 114143, 2023.

[12] L. Davoodi, J. Mezei, and M. Heikkilä, "Aspect-based sentiment classification of user reviews to understand customer satisfaction of e-commerce platforms: L. Davoodi et al.," *Electron. Commer. Res.*, vol. 26, no. 2, pp. 1417–1459, 2026.

[13] H. J. Alantari, I. S. Currim, Y. Deng, and S. Singh, "An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews," *Int. J. Res. Mark.*, vol. 39, no. 1, pp. 1–19, 2022.

[14] A. Daza, N. D. G. Rueda, M. S. A. Sánchez, W. F. R. Espíritu, and M. E. C. Quiñones, "Sentiment analysis on e-commerce product reviews using machine learning and deep learning algorithms: A bibliometric analysis, systematic literature review, challenges and future works," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 2, p. 100267, 2024.

[15] L. Ashbaugh and Y. Zhang, "A comparative study of sentiment analysis on customer reviews using machine learning and deep learning," *Computers*, vol. 13, no. 12, p. 340, 2024.

[16] Y. A. Mustofa and I. S. K. Idris, "Ensemble Approach to Sentiment Analysis of Google Play Store App Reviews," *Jambura J. Electr. Electron. Eng.*, vol. 6, no. 2, pp. 181–188, 2024.

[17] P. H. C. Samanmali and R. A. H. M. Rupasingha, "Sentiment analysis on google play store app users' reviews based on deep learning approach," *Multimed. Tools Appl.*, vol. 83, no. 36, pp. 84425–84453, 2024.

[18] C. Wang, X. Zhu, and L. Yan, "Sentiment analysis for e-commerce reviews based on deep learning hybrid model," in *Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning*, 2022, pp. 38–46.
[19] R. Kusumaningrum, I. Z. Nisa, R. Jayanto, R. P. Nawangsari, and A. Wibowo, "Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews," *Heliyon*, vol. 9, no. 6, 2023.
[20] R. Sutoyo, S. Achmad, A. Chowanda, E. W. Andangsari, and S. M. Isa, "PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks," *Data Br.*, vol. 44, p. 108554, 2022.
[21] H.-T. Duong and T.-A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput. Soc. Networks*, vol. 8, no. 1, p. 1, 2021.

.