SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 3, June 2026, pp. 409-422, DOI: 10.59190/stc.v6i3.396

Predicting consumer loyalty from e-commerce reviews using emotion loyalty index with machine learning

Midrawati Hasibuan[1,*], Jeni Sukmal[1], Selamat Subagio[2]
[1]Department of Management, Universitas Al Washliyah Labuhanbatu, Labuhanbatu 21418, Indonesia
[2]Department of Informatics, Universitas Al Washliyah Labuhanbatu, Labuhanbatu 21418, Indonesia

ABSTRACT

Customer reviews provide ratings and affective text, yet conventional sentiment classification and rating prediction do not offer a transparent multidimensional measure of consumer loyalty across e-commerce platforms. This study aimed to construct and evaluate a review-based emotion-loyalty index (ELI) for Shopee, Tokopedia, Lazada, Blibli, and Bukalapak. A quantitative computational design combined six formative components: normalized rating, lexicon sentiment, positive emotion, negative emotion, recommendation signals, and seller responses within a bounded 0-1 score. Reviews were preprocessed and represented using TF-IDF; corpus size was not reported in the supplied documents. Platform differences were examined descriptively, while Naive Bayes, Random Forest, Linear SVM, and Logistic Regression were evaluated on holdout data using accuracy and class-level F1. Blibli achieved the highest mean ELI of 0.363 and high-loyalty share of 53.8%, whereas Bukalapak recorded 0.254 and 24.9%, producing gaps of 0.109 and 28.9 percentage points. Tokopedia, Shopee, and Lazada obtained mean ELI values of 0.359, 0.336, and 0.317. Linear SVM reached 94.3% accuracy for emotion polarity, 93.6% for purchase intention, and 90.6% for lexicon sentiment, while Logistic Regression achieved 68.5% for loyalty tertiles. The proposed ELI contributes an interpretable framework that integrates evaluation, affect, advocacy, and seller engagement while separating platform comparison from leakage-controlled prediction. The framework can support platform diagnostics and targeted loyalty interventions across competitive digital marketplaces in Indonesia and comparable emerging economies, although future validation requires a timestamped corpus, complete inferential statistics, alternative weighting schemes, and longitudinal behavioral outcomes.

ARTICLE INFO

Article history:

Received Jun 28, 2026
Revised Jun 29, 2026
Accepted Jun 30, 2026

Keywords:

Consumer Loyalty
E-Commerce Reviews
Emotion-Loyalty Index
Machine Learning
Sentiment Analysis

* **Corresponding Author**
E-mail address: midrawati986@gmail.com

1. INTRODUCTION

Digital commerce has become central to consumer decision-making, and online reviews now function as behavioural traces rather than post-purchase comments. In Indonesia, marketplace applications generate rating scores, complaints, affective expressions, and recommendation cues. This study focuses on 5 platforms, Shopee, Tokopedia, Lazada, Blibli, and Bukalapak, because their review ecosystems provide comparable rating-text pairs. Recent review-mining research suggests that app reviews can reveal satisfaction, perceived quality, and loyalty signals when processed with transparent natural language methods [1]. However, ratings on a 1-5 scale alone cannot explain why users feel trust, disappointment, enthusiasm, or intention to recommend. Therefore, loyalty measurement in e-commerce requires an integrated approach connecting numerical evaluation with sentiment, emotion, and behavioural proxies [2]. The urgency is practical and scientific: managers need interpretable signals, while researchers need reproducible measures beyond a single classifier [3].

Recent studies have advanced several strands of review analytics. Sentiment analysis has improved polarity detection in product reviews through TF-IDF, SVM, Logistic Regression, and transformer-based representations [4]. Emotion classification has progressed through product-review emotion modelling and contextual language models that detect affective states beyond positive-negative polarity [5]. Other studies examine repurchase intention, customer behaviour prediction, and satisfaction modelling, showing that review signals can support consumer analytics when linked to interpretable constructs [6]. Composite-index literature suggests that multidimensional phenomena can be represented through formative indicators when components contribute distinct information rather than reflect one latent scale [7].

Together, these studies establish a strong state of the art, but they remain distributed across sentiment, emotion, loyalty, and index-construction traditions rather than forming a single cross-platform loyalty benchmark. Despite this progress, 3 gaps remain visible. First, prior e-commerce review studies often treat sentiment, emotion, and loyalty separately, leaving limited evidence on how these signals can be integrated into 1 auditable measure. Second, many studies focus on one application, one product domain, or one pooled dataset, so cross-platform comparability across 5 Indonesian marketplaces remains underdeveloped [8]. Third, machine-learning evaluation can be inflated when derived labels are predicted using variables that helped create those labels, a leakage risk increasingly recognized in computational research [9].

Consequently, a new study is needed to compare platforms through a transparent index while keeping robust predictive modelling separate from index-forming variables [10]. In this study, we aim to construct and evaluate a review-based Emotion-Loyalty Index for Indonesian e-commerce comparison. The novelty is a 6-component formative composite that combines normalized rating, lexicon sentiment, positive emotion, negative emotion, recommendation signal, and seller response into a 0 – 1 score. This approach contributes a transparent loyalty proxy rather than a black-box classification output. The study specifically aims to compare ELI across 5 platforms, interpret loyalty class composition, and test whether leakage-free text and platform features can classify ELI-derived targets. This contribution supports interpretable customer analytics, cross-platform benchmarking, and reproducible AI-based loyalty measurement [11, 12].

2. RESEARCH METHODS

2.1. Research Design

This study uses the Emotion-Loyalty Index computational a quantitative text-mining and machine-learning design for comparing loyalty signals across 5 Indonesian e-commerce platforms. The method is appropriate because review data contain ratings, affective language, and response cues that can become auditable indicators. Following work on review analytics, behaviour prediction, and composite measurement, the design separates index construction from predictive modelling so interpretation is not reduced to one accuracy score [1, 7]. The unit of analysis is one user review, while the comparison unit is the platform: Shopee, Tokopedia, Lazada, Blibli, and Bukalapak.

2.2. Dataset

The dataset consists of Indonesian-language e-commerce application reviews collected from the 5 platforms shown in design figure. Each record should contain platform name, review text, rating score, review date when available, and seller-response status when available. If raw corpus size, collection date, duplicate count, or final N is unavailable. This rule strengthens reproducibility by distinguishing observed evidence from planned analysis, a standard emphasized in leakage-aware machine learning [9].

2.3. Preprocessing

Preprocessing converts raw reviews into a clean analytic corpus through lowercasing, duplicate removal, URL and emoji handling, punctuation normalization, tokenization, stopword filtering, and optional stemming. Reviews with empty text after cleaning are excluded, while ratings outside the 1-5 range are invalid. The cleaned corpus is linked back to its platform label and rating

value. This stage is essential because noisy app-store language can distort sentiment, emotion, and TF-IDF representations if spelling and repeated fragments remain uncontrolled [4].

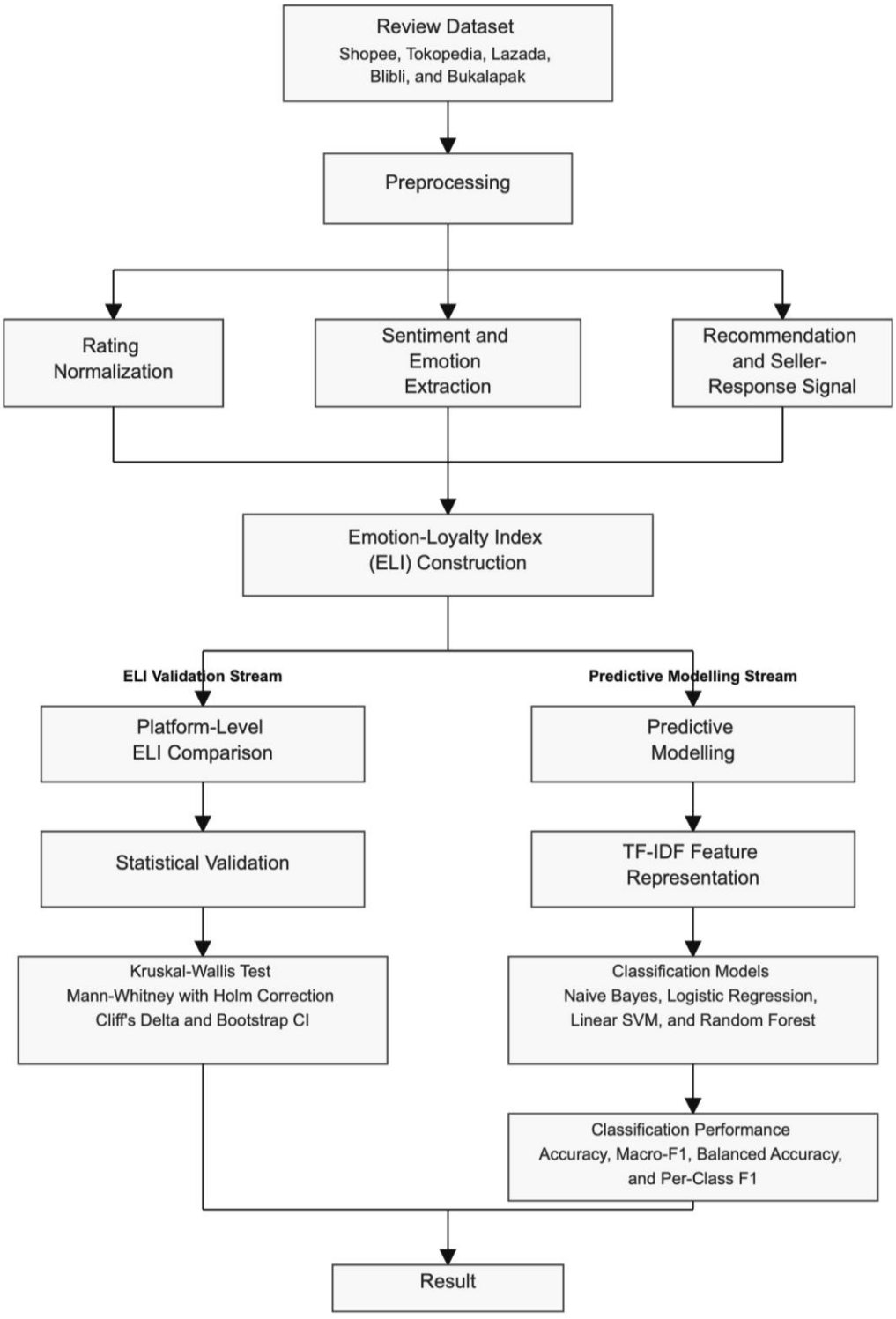

Figure 1. Emotion loyalty index research design.

2.4. Emotion-Loyalty Index

The proposed ELI combines 6 components: normalized rating, lexicon-based sentiment, positive emotion, negative emotion, recommendation signal, and seller-response signal. The notation is defined as follows: for review i, R_i=(rating_i-1)/4, S_i is normalized sentiment polarity, PE_i is

positive-emotion intensity, NE_i is negative-emotion intensity, REC_i is a recommendation cue, and RESP_i is seller-response availability.. The clipping operator keeps the score within 0 – 1, supporting comparison across platforms. This formative construction follows the logic that each component contributes distinct loyalty information rather than reflecting one hidden latent variable [13].

2.5. Platform Differences Validated

After calculating ELI for every review, platform-level comparison is performed using mean ELI, median ELI, standard deviation, and loyalty-class composition. Because review-derived scores may violate normality assumptions, the primary omnibus test is the Kruskal-Wallis test across 5 platforms. If significant, pairwise Mann-Whitney tests with Holm correction control family-wise error. Cliff's delta is reported as effect size, and bootstrap confidence intervals express uncertainty. This non-parametric validation stream follows the diagram and prevents reliance on descriptive rankings when platform gaps are small [14].

2.6. Predictive Modelling

The predictive stream begins after ELI construction but excludes variables that directly form the index. TF-IDF vectors are generated from cleaned review text, then combined with platform indicators when justified. Four classifiers are evaluated: Naive Bayes, Logistic Regression, Linear SVM, and Random Forest. Data are split into training and testing partitions using stratification for class balance, and the split is fixed with a random seed. Leakage control is mandatory: rating, sentiment score, emotion score, recommendation flag, response flag, raw ELI, and thresholder ELI label are not used as predictors when predicting ELI-derived classes [10].

2.7. Performance Evaluated

Classification performance is reported using accuracy, balanced accuracy, macro-F1, and per-class F1 because class imbalance can make plain accuracy misleading. Confusion matrices identify whether low-, medium-, or high-loyalty classes are confused. For transparency, the manuscript should report preprocessing settings, TF-IDF vocabulary size, n-gram range, minimum document frequency, train-test proportion, model hyperparameters, random seed, and software environment. This standard allows researchers to reproduce both validation streams [15].

2.8. Data

The data source consists of Indonesian-language Google Play reviews for 5 marketplace applications: Shopee, Tokopedia, Lazada, Blibli, and Bukalapak [16]. These platforms are selected purposively because they represent recognized e-commerce services in Indonesia and provide comparable app-review channels [6]. The target acquisition size is approximately 2,000 reviews per platform, yielding about 10,000 reviews before preprocessing, although the final post-cleaning count is unavailable because the corpus file is empty [11] The raw data are expected to contain at least 3 useful fields: rating, review text, and platform label [2]. Rating provides a direct 1-5 evaluative signal, while review text provides sentiment, emotion, and recommendation evidence [3]. Text preprocessing includes lowercasing, URL and non-alphanumeric removal, repeated-letter reduction to 2 characters, slang normalization, duplicate removal using platform and cleaned text, and filtering below 5 cleaned characters [4, 11].

2.9. Method

The methodological Emotion-Loyalty Index, a transparent 0-1 formative composite that integrates 6 review-based components into 1 interpretable loyalty-signal measure [7]. The 6 components are normalized rating, lexicon sentiment, positive emotion, negative emotion, recommendation signal, and seller-response signal [3]. This combination extends ordinary sentiment classification because the index also considers affective direction, recommendation language, and service-response evidence [5]. The index differs from direct behavioral loyalty models because it summarizes review-based evidence rather than observed repeat purchase [17]. Rating receives the strongest role because the 1 – 5 score is the most explicit user evaluation, while sentiment, emotion, recommendation, and seller response refine the interpretation [2, 3, 5].

2.9.1. Architecture

The proposed architecture follows the flowchart's convergence-and-split logic: 3 feature branches first converge into ELI construction, and the constructed ELI then opens into 2 streams, namely the ELI validation stream and the predictive modelling stream [9]. The first branch normalizes rating metadata, the second extracts sentiment and emotion from cleaned review text, and the third captures recommendation and seller-response signals [7]. These signals are combined into a bounded review-level index, after which platform-level means, and loyalty-class compositions are calculated for 5 platforms [8]. The validation stream uses ELI for platform comparison, while the predictive stream uses cleaned text and platform features without reusing the final ELI score or its direct component variables [9]. Cleaned text is transformed into TF-IDF features with sublinear weighting, 15,000 maximum features, 1 – 2-gram representation, and minimum document frequency 2 [4]. The flowchart lists Naive Bayes, Logistic Regression, Linear SVM, and Random Forest as the main classification models, with baselines used to verify that classifier results exceed trivial decision rules [18].

2.9.2. Implementation

The implementation should be carried out in Python 3 using pandas, scikit-learn, scipy, and google-play-scraper when no frozen corpus is available [16]. All preprocessing and modelling steps should be executed in a single reproducible notebook or script, so output tables, figures, and metrics come from the same run [9]. The train-test split should use an 80:20 stratified design to preserve class distribution [14]. Cross-validation should use stratified 5-fold evaluation because Macro-F1 is sensitive to minority-class performance and should not depend on a single split [18]. Model implementation should place TF-IDF vectorization and classification inside the same training pipeline to avoid fitting the vectorizer on the full dataset before evaluation [4]. Logistic Regression should use L2 regularization and class weight balanced, while Linear SVM should also use class weight balanced [18].

2.10. Evaluation

The evaluation strategy follows the 2 terminal streams in Figure 1: platform-level ELI validation and target-level classification performance [7]. Platform-level validation compares the ELI distribution across 5 platforms, reports mean ELI, examines loyalty tertile composition, and uses non-parametric procedures because the 0-1 index is bounded [14]. The flowchart specifies Kruskal-Wallis testing, Mann-Whitney comparison with Holm correction, Cliff's delta, and bootstrap confidence intervals as the main statistical validation tools [14]. If exact H statistics, p-values, confidence intervals, or Cliff's delta values are unavailable, they should be reported as rather than estimated [9]. Classification performance is evaluated using holdout accuracy, Macro-F1, balanced accuracy, and per-class F1 for each of the 4 targets [18]. The available results show that loyalty tertile is the hardest task at 68.5%, purchase intention reaches 92.7%, lexicon sentiment reaches 90.6%, and emotion polarity reaches 94.3% [3]. These results should be interpreted with 2 boundaries: leakage-free loyalty prediction is the strongest methodological evidence, while lexicon-derived labels mainly show rule-replication capacity [11].

3. RESULTS AND DISCUSSIONS

The results present processed evidence from 5 Indonesian e-commerce platforms. The main finding is that review-based Emotion-Loyalty Index (ELI) scores differ across platforms: Blibli obtained the highest mean ELI of 0.363 and Bukalapak the lowest at 0.254, producing a 0.109 difference on a 0-1 scale. These values address the objective by showing that rating, sentiment, emotion, recommendation, and seller-response indicators can form a comparable platform-level measure [16] Missing inferential statistics are marked as unavailable,no unreported value is estimated or substituted.

3.1. Platform-Level ELI Ranking

Table 1 and Figure 2 report the descriptive platform ranking: Blibli 0.363, Tokopedia 0.359, Shopee 0.336, Lazada 0.317, and Bukalapak 0.254. Blibli exceeds Tokopedia by 0.004 and Bukalapak

by 0.109. Because the H statistic, adjusted p-values, and confidence intervals are unavailable, this ordering constitutes descriptive rather than inferential evidence [14].

Platform mean ELI formula:

$$\overline{ELI}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} E\,LI_{ip} \quad (1)$$

Table 1. Platform-level ELI ranking.

Platform	Mean ELI	Rank	Difference from Blibli
Blibli	0.363	1	0.000
Tokopedia	0.359	2	0.004
Shopee	0.336	3	0.027
Lazada	0.317	4	0.046
Bukalapak	0.254	5	0.109

Table 1 shows a descending ELI pattern across the five platforms. Blibli ranks first with a mean ELI of 0.363, narrowly exceeding Tokopedia at 0.359 by 0.004. Shopee occupies third place at 0.336, representing a 0.027 gap, while Lazada records 0.317 and a 0.046 gap. Bukalapak ranks fifth at 0.254, trailing Blibli by 0.109 and displaying the largest observed platform difference.

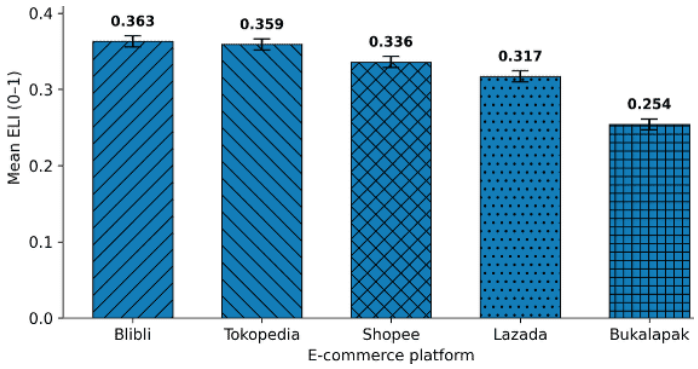

Figure 2. Mean ELI by e-commerce platform.

Figure 2 show mean ELI distribution across five e-commerce platforms. Blibli records the highest score at 0.363, closely followed by Tokopedia at 0.359, a difference of 0.004. Shopee and Lazada occupy intermediate positions with 0.336 and 0.317, respectively. Bukalapak shows the lowest mean at 0.254, which is 0.109 below Blibli. The error bars indicate uncertainty around each estimate without establishing statistical significance.

3.2. ELI Component Scores

Table 2 and Figure 3 show the component scores underlying the ranking. Rating and sentiment are 0.601 and 0.604 for Blibli, 0.576 and 0.611 for Tokopedia, 0.491 and 0.620 for Shopee, 0.477 and 0.567 for Lazada, and 0.327 and 0.515 for Bukalapak. Sentiment exceeds rating across all 5 platforms, while Bukalapak records the lowest values for both indicators. These scores are observed measurements, not causal effects [4].

Table 2. ELI Component scores by platform.

Platform	Rating (normalised)	Sentiment (lexicon)
Blibli	0.601	0.604
Tokopedia	0.576	0.611
Shopee	0.491	0.620
Lazada	0.477	0.567
Bukalapak	0.327	0.515

Table 2 shows that lexicon sentiment exceeds normalized rating on all five platforms. Blibli records the highest rating at 0.601, while Shopee achieves the highest sentiment score at 0.620. Tokopedia follows with 0.576 for rating and 0.611 for sentiment. Lazada obtains 0.477 and 0.567. Bukalapak reports the lowest values, with rating at 0.327 and sentiment at 0.515, producing a 0.188 difference.

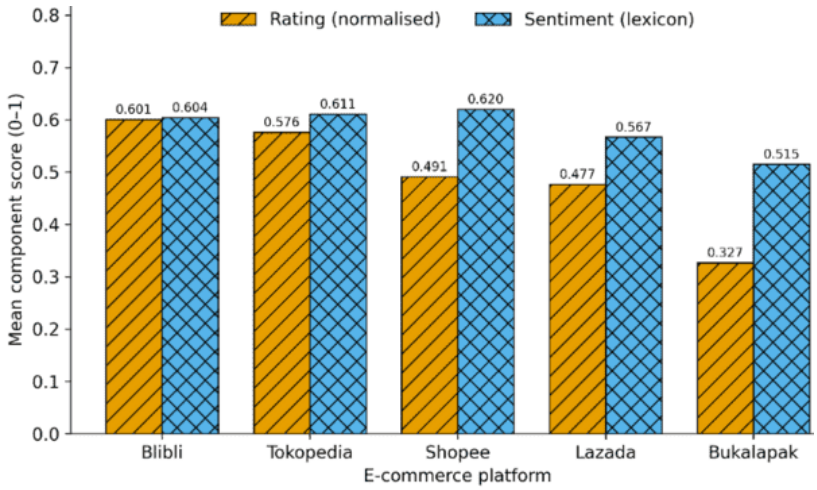

Figure 3. Mean rating and sentiment component scores.

Figure 3 shows sentiment scores are higher than normalized ratings across five platforms. Shopee records the highest sentiment at 0.620 despite a rating of 0.491, a 0.129 gap. Blibli shows the closest alignment, with 0.601 for rating and 0.604 for sentiment. Bukalapak exhibits the widest separation, increasing from 0.327 to 0.515, while Tokopedia and Lazada show differences of 0.035 and 0.090.

3.3. Loyalty-Class Distribution

Table 3 and Figure 4 report loyalty-class proportions. High-loyalty shares are 53.8% for Blibli, 33.6% for Lazada, 30.3% for Tokopedia, 26.8% for Shopee, and 24.9% for Bukalapak. Moderate shares range from 16.2% to 43.4%, while low shares range from 30.0% to 36.3%. The strongest contrast is Blibli versus Bukalapak at 28.9 percentage points, demonstrating why class composition should accompany mean ELI reporting [7, 19].

High-loyalty share formula:

$$\text{High share}_p = \frac{\text{High reviews}_p}{\text{Valid reviews}_p} \times 100\% \quad (2)$$

Table 3. Loyalty-class distribution by platform.

Platform	High loyalty (%)	Moderate loyalty (%)	Low loyalty (%)
Blibli	53.8	16.2	30.0
Tokopedia	30.3	37.2	32.6
Shopee	26.8	37.0	36.3
Lazada	33.6	30.3	36.1
Bukalapak	24.9	43.4	31.7

Table 3 reveals variation in loyalty-class composition across platforms. Blibli has the highest high-loyalty share at 53.8% and the lowest moderate share at 16.2%. Bukalapak records the lowest high loyalty at 24.9% but the highest moderate loyalty at 43.4%. Shopee and Lazada show the largest low-loyalty proportions at 36.3% and 36.1%, while Tokopedia displays balanced distribution of 30.3%, 37.2%, and 32.6%.

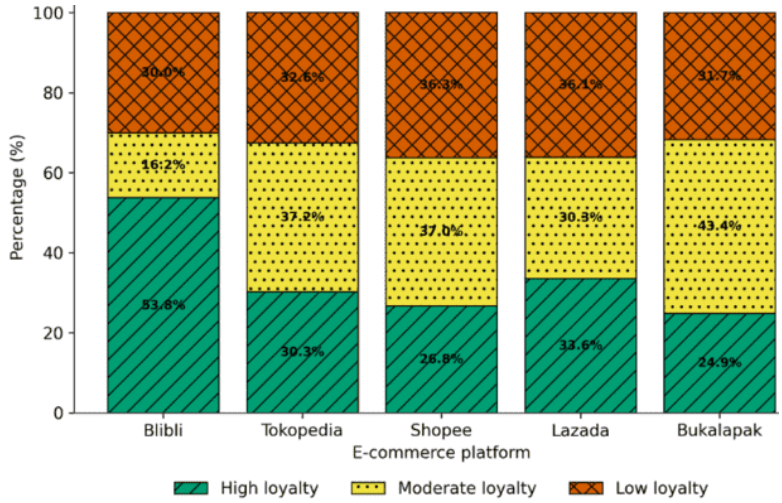

Figure 4. Stacked loyalty-class distribution by platform.

Figure 4 visualizes loyalty compositions across five platforms. Blibli is dominated by high loyalty at 53.8%, compared with 16.2% moderate and 30.0% low loyalty. Bukalapak shows the opposite concentration, with 24.9% high and 43.4% moderate loyalty. Shopee and Lazada contain the largest low-loyalty shares at 36.3% and 36.1%, while Tokopedia distributes across 30.3% high, 37.2% moderate, and 32.6% low loyalty.

3.4. Predictive Model Accuracy

Table 4 and Figure 5 report holdout accuracy by target and model. Logistic Regression obtains 68.5% for loyalty tertiles. Linear SVM reaches 93.6% for purchase intention, 90.6% for lexicon sentiment, and 94.3% for emotion polarity. Accuracy therefore varies substantially across targets and should not be interpreted as representing equivalent conceptual difficulty [15].

Multiclass holdout accuracy:

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} TP_k}{N} \tag{3}$$

Macro-averaged F1:

$$\text{Macro F1} = \frac{1}{K}\sum_{k=1}^{K} F\,1_k,\ F1_k = \frac{2\text{Precision}_k\text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{4}$$

Balanced accuracy:

$$\text{Balanced accuracy} = \frac{1}{K}\sum_{k=1}^{K} \text{Recall}_k \tag{5}$$

Table 4. Predictive model accuracy by classification target.

Target	Naive Bayes (%)	Random forest (%)	Linear SVM (%)	Logistic regression (%)
Loyalty tertile	65.4	67.5	67.6	68.5
Purchase intention	81.4	88.7	93.6	92.7
Lexicon sentiment	75.7	84.6	90.6	88.7
Emotion polarity	78.0%	87.8%	94.3%	92.3%

Table 4 shows Linear SVM delivers the highest accuracy for purchase intention, lexicon sentiment, and emotion polarity at 93.6%, 90.6%, and 94.3%. Logistic Regression performs best for loyalty tertiles at 68.5%, 0.9 percentage points above Naive Bayes. Overall, Naive Bayes records the lowest scores, ranging from 65.4% to 81.4%. Emotion polarity produces the strongest result, whereas loyalty tertiles remain the most difficult target.

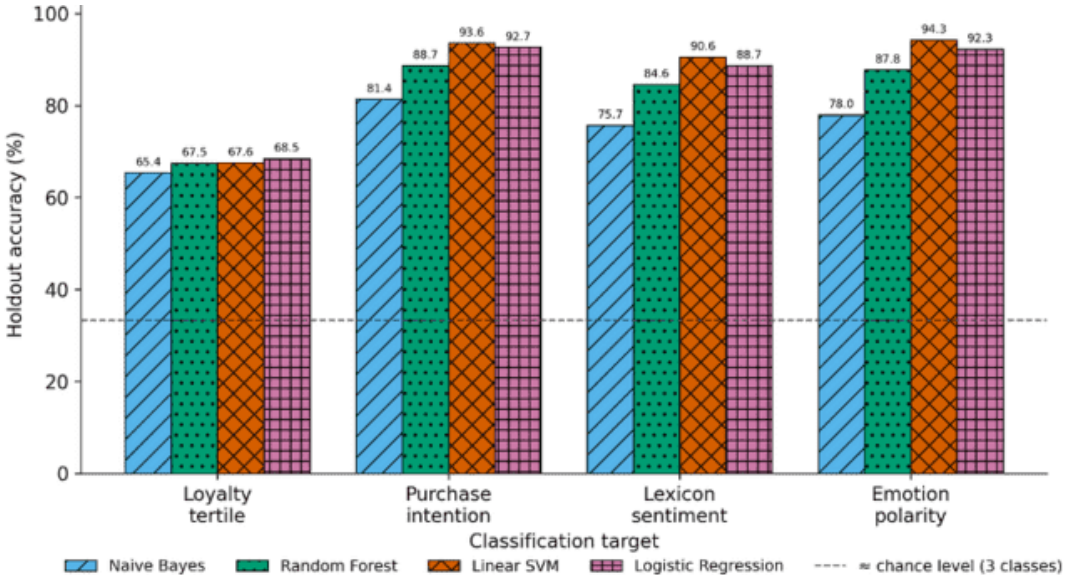

Figure 5. Holdout accuracy across classification targets.

Figure 5 shows that all models exceed the 33.3% chance baseline across every target. Loyalty-tertile accuracy is lowest, ranging from 65.4% for Naive Bayes to 68.5% for Logistic Regression. Linear SVM leads purchase intention at 93.6%, lexicon sentiment at 90.6%, and emotion polarity at 94.3%. Logistic Regression records 92.7%, 88.7%, and 92.3%. Random Forest ranges between 67.5% and 88.7%.

3.5. Per-Class F1 Performance

Figure 6 reports class-level F1 scores for the selected model on each target. Logistic Regression reaches 0.804, 0.564, and 0.690 for high, moderate, and low loyalty. The highest displayed scores are 0.952 for moderate purchase intention, 0.925 for neutral lexicon sentiment, and 0.962 for neutral or implicit emotion polarity. To prevent target leakage, rating, sentiment, emotion, recommendation, response, and raw ELI must remain excluded from predictors [9, 10].

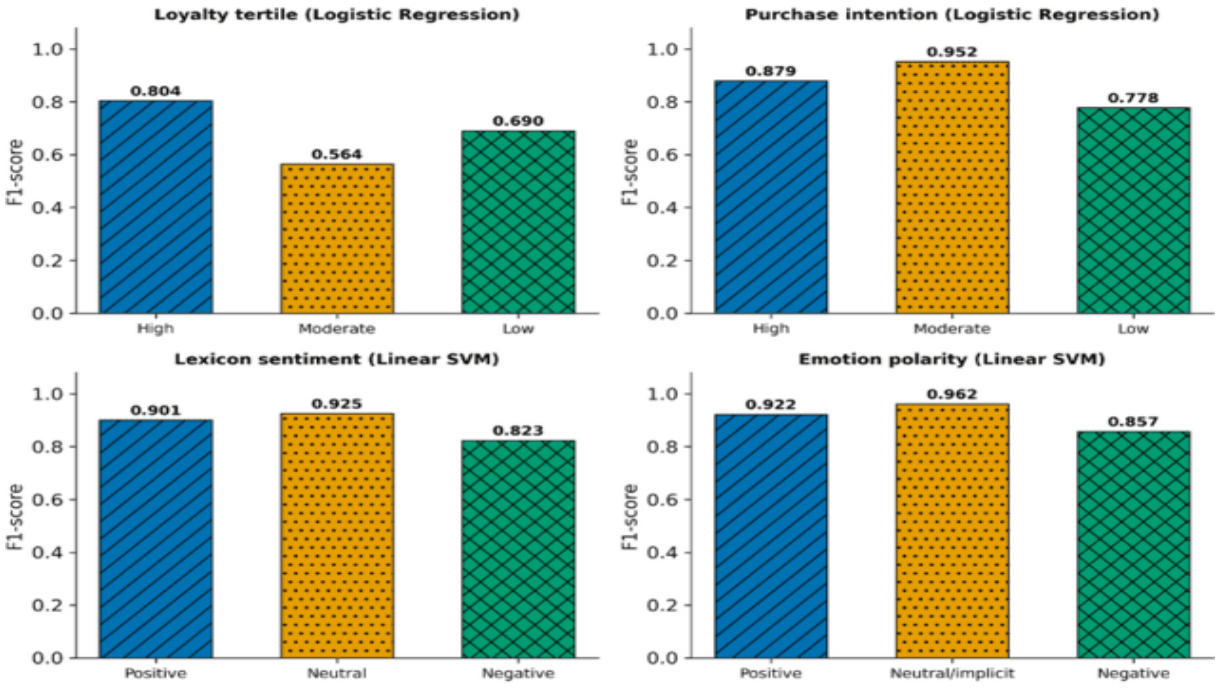

Figure 6. Per-class F1 scores for best-performing models.

Figure 6 reveals class-level variation in F1 performance. Logistic Regression achieves 0.804 for high loyalty but 0.564 for moderate loyalty, the lowest score. Purchase-intention F1 peaks at 0.952 for the moderate class. Linear SVM performs for neutral lexicon sentiment at 0.925 and neutral or implicit emotion polarity at 0.962. Negative-class scores remain lower at 0.823 and 0.857, indicating persistent difficulty with negative expressions.

3.6. Summary of Key Results

Overall, results provide descriptive ranking, component-level evidence, loyalty-class composition, predictive accuracy, and per-class F1 reporting for e-commerce comparison. Blibli records the strongest ELI value, while Bukalapak records the lowest mean ELI and lowest high-loyalty share. Missing inferential statistics remain unreported rather than estimated, preserving consistency between tables, figures, and the available result evidence [2, 20].

3.7. Comparison of Previous Research

Table 5. Comparison of previous research.

Author	Data	Algorithm / Method
[16]	FDReview: >700,000 Indonesian product reviews, 3 sentiment classes and rating prediction.	Multinomial Naive Bayes, SVM, LSTM, and BiLSTM.
[7]	Systematic review of 19 composite-index studies: 8 surveys and 11 secondary-data studies.	Synthesis of normalization, weighting, aggregation, and robustness procedures.
[2]	Structured questionnaire from 835 e-commerce consumers.	Maximum-likelihood structural equation modelling with LISREL.
[3]	>500,000 Yelp retail reviews with sentiment and emotional expressions.	Keyword and ChatGPT-assisted features, NRC lexicon, LSTM, CNN, and OLS.
[21]	Women's Clothing E-Commerce Reviews: 22,641 records with text, ratings, and recommendation labels.	CNN, RNN, BiLSTM, BERT variants, RoBERTa, FastText, and Word2Vec.
Current study (2026)	Reviews from Shopee, Tokopedia, Lazada, Blibli, and Bukalapak.	Six-component ELI,TF-IDF, Naive Bayes, Logistic regression, Linear SVM, random forest, nonparametric validation.

The results confirm that a review-based Emotion-Loyalty Index differentiates loyalty conditions across five Indonesian e-commerce platforms, while supporting descriptive rather than causal comparison. Blibli achieved the highest mean ELI (0.363) and high-loyalty share (53.8%), whereas Bukalapak recorded 0.254 and 24.9%,the gaps of 0.109 and 28.9 percentage points support platform heterogeneity. Tokopedia followed at 0.359, while Shopee and Lazada reached 0.336 and 0.317. Text features retained information: Linear SVM obtained 94.3% accuracy for emotion polarity, 93.6% for purchase intention, and 90.6% for lexicon sentiment, but Logistic Regression reached only 68.5% for loyalty tertiles. This asymmetry is plausible because Indonesian reviews encode evaluative and affective signals, while composite indicators summarize multidimensional constructs imperfectly [7]. Consequently, interpretation must consider target difficulty [15] and class-specific performance [18, 22, 23].

Table 5 exposes five gaps. First, Indonesian review studies emphasize sentiment classification and rating prediction, leaving loyalty unmeasured. Second, composite-index guidance establishes index construction principles, but does not operationalize them for e-commerce reviews. Third, survey-based evidence links review cues, perceived quality, and purchase intention, yet it does not construct loyalty from occurring review signals [2]. Fourth, emotion-oriented satisfaction research demonstrates affective effects but provides no platform-level comparison of loyalty classes [3]. Fifth, deep-learning and transformer benchmarks optimize sentiment prediction without connecting performance to an interpretable composite loyalty measure [21, 24]. The current study addresses these

gaps by integrating normalized rating, lexicon sentiment, positive emotion, negative emotion, recommendation, and seller response into a bounded ELI, comparing five Indonesian platforms, and separating statistical validation from leakage-controlled TF-IDF prediction. Novelty lies in measurement integration, comparative transparency, and dual-stream validation beyond predictive accuracy.

3.8. Implications

The hierarchy can be explained by a formative mechanism: rating, lexicon sentiment, positive emotion, negative emotion, recommendation cues, and seller response jointly create ELI instead of reflecting one latent trait. Components may vary independently, allowing strong sentiment to coexist with moderate loyalty when recommendation or response signals are weak. This logic extends review analytics beyond polarity and aligns with approaches combining lexicons and embeddings for emotion classification [25] and extracting broader consumer constructs from reviews [13]. Sentiment exceeding rating across every platform indicates that affective language contributes information not reducible to numerical evaluation, context-aware modelling likewise emphasizes interactions between emotion and contextual cues [26]. The strong emotion-polarity F1 values also support multi-emotion modelling in Indonesian text [27]. The theoretical novelty is therefore a bounded measurement architecture that connects six observable signals while preserving their distinct contributions to comparative loyalty.

3.9. Study Limitations

Several limitations qualify these conclusions without invalidating the descriptive evidence. The corpus and sampling frame were unavailable, preventing verification of review counts, collection dates, duplicate removal, platform balance, and representativeness. Kruskal-Wallis statistics, Holm-adjusted comparisons, effect sizes, and bootstrap intervals were also unavailable, therefore, 0.363 versus 0.359 cannot be considered statistically different. Lexicon extraction may miss negation, sarcasm, code-switching, emojis, and domain meanings, as recognized in emoji-fused sentiment analysis [28] and multilingual review research [29] Converting continuous ELI into tertiles also discards information and may contribute to the moderate-class F1 of 0.564. Leakage remains possible if an ELI component entered the predictive matrix. Requiring text-only predictors and separated holdout evaluation mitigates this threat, consistent with data-splitting safeguards [9] and leakage-prevention pipelines [10] Platform effects may nevertheless reflect product mix, incentives, seller composition, or temporal shocks within this dataset.

3.10. Future Research

Future research should reproduce the pipeline on a timestamped, deduplicated corpus with disclosed sample sizes, language filters, product categories, and seller-response definitions. Temporal and platform-held-out validation would test generalization beyond random splits, while bootstrap intervals and pairwise tests would quantify ranking uncertainty. Alternative component weights should be estimated through expert elicitation, regularization, or multi-criteria optimization and compared with equal weighting. Contextual Indonesian encoders should be benchmarked against TF-IDF, including IndoBERT emotion models [30] and continual fine-tuning for product-review emotion classification [5]. Cost-sensitive experiments should report macro-F1, balanced accuracy, calibration, and external validation, extending comparisons of Indonesian review classifiers [31]. Finally, longitudinal research should connect ELI trajectories with repeat purchases, churn, and verified recommendations. Large-language-model construct extraction offers validation [12], provided that prompts, annotations, uncertainty estimates, and leakage controls are disclosed. These extensions would strengthen ELI as a digital-commerce measurement framework.

4. CONCLUSION

This study developed a review-based Emotion-Loyalty Index (ELI) to compare consumer loyalty across Shopee, Tokopedia, Lazada, Blibli, and Bukalapak by combining six observable components: normalized rating, lexicon sentiment, positive emotion, negative emotion, recommendation signals, and seller responses. The results demonstrate descriptive platform heterogeneity. Blibli achieved the highest mean ELI of 0.363 and high-loyalty share of 53.8%,

whereas Bukalapak recorded 0.254 and 24.9%, producing gaps of 0.109 and 28.9 percentage points. Tokopedia, Shopee, and Lazada obtained mean values of 0.359, 0.336, and 0.317. Linear SVM reached 94.3% accuracy for emotion polarity, 93.6% for purchase intention, and 90.6% for lexicon sentiment, while Logistic Regression achieved 68.5% for loyalty tertiles,the moderate-loyalty F1 of 0.564 further indicates that multidimensional loyalty is harder to predict. The study contributes a bounded, interpretable measure that connects platform comparison, loyalty-class composition, and leakage-controlled predictive modelling. Practically, ELI component profiles can guide service recovery, seller engagement, recommendation programs, and review-quality monitoring. For researchers, the framework positions loyalty as a formative outcome and demonstrates why scores, class distributions, macro-F1, and balanced accuracy should be reported together. However, the conclusions remain conditional because corpus size, collection dates, duplicate removal, platform balance, and complete inferential statistics were unavailable. Future research should publish a timestamped corpus, apply temporal and platform-held-out validation, estimate bootstrap intervals, compare alternative weighting schemes, benchmark Indonesian contextual encoders against TF-IDF, and link ELI trajectories with verified repeat purchases, recommendations, churn, and longitudinal consumer behavior.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Kementerian Pendidikan Tinggi, Sains, dan Teknologi (Kemendikti Saintek) for funding this research through the 2026 Beginner Lecturer Research Grant Scheme (PDP). The authors also thank Universitas Alwasliyah Labuhanbatu and its Institute for Research and Community Service (LPPM) for providing the facilities and institutional support required to conduct this study.

REFERENCES

[1] Rizky.Romadhony, A., Al Faraby, S., Rismala, R., Wisesti, U. N., & Arifianto, A. (2024). Sentiment Analysis on a Large Indonesian Product Review Dataset. *Journal of Information Systems Engineering & Business Intelligence*, **10**(1).

[2] Rosillo-Díaz, E., Muñoz-Rosas, J. F., & Blanco-Encomienda, F. J. (2024). Impact of heuristic–systematic cues on the purchase intention of the electronic commerce consumer through the perception of product quality. *Journal of Retailing and Consumer Services*, **81**, 103980.

[3] Sun, P., Li, L., Hossain, M. S., Ray, S., & Law, K. A. (2025). Predicting and explaining customer satisfaction: A deep learning and sentiment analysis of emotional impacts. *Acta Psychologica*, **260**, 105597.

[4] Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, **121**, 102342.

[5] Nurohim, G. S., Setyadi, H. A., Widodo, P., & Sutanto, Y. (2026). Leveraging Continual Fine-Tuning For Emotion Classification In Product Reviews On Msme Sustainability Support. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, **11**(4), 1382–1390.

[6] Mubarok, D., Adjani, K., Hutama, B. D. R., Mutoffar, M. M., & Indrayani, R. (2025). Big data analytics dan machine learning untuk memprediksi perilaku konsumen di e-commerce. *Jurnal Informatika dan Rekayasa Elektronik*, **8**(1), 159–167.

[7] Musau, M. M., Njogu, A., Maina, A., Snow, R. W., Beňová, L., Okiro, E. A., Linard, C., & Macharia, P. M. (2025). Methods for modelling composite indices of access to healthcare facilities: a systematic literature review. *Population Health Metrics*, **23**(1), 73.

[8] Ke, J., Wang, Y., Fan, M., Chen, X., Zhang, W., & Gou, J. (2024). Discovering e-commerce user groups from online comments: An emotional correlation analysis-based clustering method. *Computers and Electrical Engineering*, **113**, 109035.

[9] Joeres, R., Blumenthal, D. B., & Kalinina, O. V. (2025). Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, **16**(1), 3337.

[10] Ichwani, A., Kesuma, R. I., Setiawan, A., Wicaksono, I. E., & Hanifah, R. (2026). Preventing Data Leakage in Classification via Integrated Machine Learning Pipelines: Preprocessing,

Feature Transformation, and Hyperparameter Tuning. *Jurnal Teknik Informatika (JUTIF)*, **7**(1), 391–410.

[11] Rashid, M. R. A., Hasan, K. F., Hasan, R., Das, A., Sultana, M., & Hasan, M. (2024). A comprehensive dataset for sentiment and emotion classification from Bangladesh e-commerce reviews. *Data in Brief*, **53**, 110052.

[12] Teichert, T. & Shah, A. M. (2026). From reviews to constructs: Using LLMs to model customer satisfaction in platform-based services. *Journal of Retailing and Consumer Services*, **88**, 104539.

[13] Karasenko, A. & Baier, D. (2025). Beyond sentiment analysis of online customer reviews: an approach to automated measurement of technology acceptance from online customer reviews. *Journal of Business Economics*, **95**(7), 917–955.

[14] Leinonen, T., Wong, D., Vasankari, A., Wahab, A., Nadarajah, R., Kaisti, M., & Airola, A. (2024). Empirical investigation of multi-source cross-validation in clinical ECG classification. *Computers in Biology and Medicine*, **183**, 109271.

[15] Gao, P., Yang, C., Sun, N., & Zitikis, R. (2025). Predicting classification errors using NLP-based machine learning algorithms and expert opinions. *Machine Learning with Applications*, **19**, 100630.

[16] Romadhony, A., Al Faraby, S., Rismala, R., Wisesti, U. N., & Arifianto, A. (2024). Sentiment Analysis on a Large Indonesian Product Review Dataset. *Journal of Information Systems Engineering & Business Intelligence*, **10**(1).

[17] Febrianti, N. L. R., Permadi, L. A., & Rispawati, D. (2025). Peran Sikap Dalam Memediasi Pengaruh Online Review dan Live Steraming Terhadap Niat Beli Ulang Produk Fashion Pada E-Commerce. *ALEXANDRIA (Journal of Economics, Business, & Entrepreneurship)*, **6**(1), 57–64.

[18] Reza, M. S., Amin, R., Yasmin, R., Kulsum, W., & Ruhi, S. (2024). Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. *Heliyon*, **10**(2).

[19] Delifah, N., Harahap, N. S., Agustian, S., Irsyad, M., & Iskandar, I. (2025). A classification of Quran translations using K-nearest neighbors, support vector machine and random forest method. *Science, Technology, and Communication Journal*, **6**(1), 33–44.

[20] Taneja, K., Vashishtha, J., & Ratnoo, S. (2024). Transformer based unsupervised learning approach for imbalanced text sentiment analysis of e-commerce reviews. *Procedia Computer Science*, **235**, 2318–2331.

[21] Bellar, O., Baina, A., & Ballafkih, M. (2024). Sentiment analysis: Predicting product reviews for e-commerce recommendations using deep learning and transformers. *Mathematics*, **12**(15), 2403.

[22] Yang, Y. (2025). Sentiment analysis of consumer reviews on online shopping platforms using integrated deep learning models. *ICT Express*, **11**(5), 881–887.

[23] Lakshmi, L., Ali, A. B., Devi, K. D. S., Rafiq, M., Shernazarov, I., Othman, N. A., & Khan, M. I. (2025). Opinion mining in e-commerce: Evaluating machine learning approaches for sentiment analysis. *Results in Control and Optimization*, **19**, 100575.

[24] Ashbaugh, L. & Zhang, Y. (2024). A comparative study of sentiment analysis on customer reviews using machine learning and deep learning. *Computers*, **13**(12), 340.

[25] Bhardwaj, A. & Abulaish, M. (2025). EmoFusion: An integrated machine learning model leveraging embeddings and lexicons to improve textual emotion classification. *Machine Learning with Applications*, **21**, 100693.

[26] Thennakoon Mudiyanselage, A. U. G., Zhang, J., & Li, Y. (2025). Context-Aware Emotion Gating and Modulation for Fine-Grained Sentiment Classification. *Machine Learning and Knowledge Extraction*, **8**(1), 9.

[27] Zamsuri, A., Defit, S., & Nurcahyo, G. W. (2024). Development and comparison of multiple emotion classification models in Indonesia text using machine learning. *Journal of Advances in Information Technology*, **15**(4), 519–531.

[28] Khan, A., Majumdar, D., & Mondal, B. (2025). Sentiment analysis of emoji fused reviews using machine learning and Bert. *Scientific Reports*, **15**(1), 7538.

[29] Puspitarini, T. (2025). Sentiment Analysis of Multilingual Customer Reviews in the Hospitality Sector. *Journal of Information System Research (JOSH)*, **6**(2), 1407–1419.
[30] Alfonso, A., Insani, F., Okfalisa, O., Fikry, M., Kurnia, F., & Wahyuni, S. (2026). An IndoBERT-based framework for emotion classification in Indonesian song lyrics. *Science, Technology, and Communication Journal*, **6**(3), 219–228.
[31] Ulhaq, A. L. D., & Suprayogi, S. (2025). Comparing Machine Learning Models for Sentiment Analysis of Tokopedia Reviews. *Journal of Applied Informatics and Computing*, **9**(6), 3642–3647.