Popularity Bias and Category Diversity in Review-Based Rating Prediction for E-Commerce: An Empirical Study on Indonesian Marketplace Data

Andy Supriyadi[1]*, Herman Dwi Surjono[2], Handaru Jati[3]
Doctoral Program in Engineering Science-Faculty of Engineering, Universitas Negeri Yogyakarta, Indonesia[1, 2, 3]
Department of Informatics Engineering-Vocational School, Universitas Sebelas Maret Surakarta, Indonesia[1]

***Abstract*—Online marketplaces increasingly rely on user reviews to estimate product quality. However, most empirical comparisons of review-based prediction models report only aggregate accuracy and overlook how item popularity and category composition shape that accuracy. This study presents a controlled empirical analysis of review-based rating prediction on a self-collected, anonymized corpus of 44,229 Indonesian marketplace reviews spanning four product categories (fashion, electronics, tools/hardware, and sports). To isolate the effect of category diversity from data volume, the experimental design was set with the total number of reviews kept constant at 12,000, while the number of categories was increased from two to four. For product metadata classification, three text classification models (TF-IDF with logistic regression, linear SVM, and random forest) and a customized IndoBERT transformer model were compared using five-fold cross-validation, and the difficulty of prediction was further analyzed at the item level. Three findings emerge. First, the review text is the dominant signal, reducing the mean absolute error to 0.62 (0.57 with IndoBERT), compared with 1.28 for the majority baseline. Second, increasing category diversity within a fixed budget results in a small but consistent performance degradation. Third, and most notably, popular items are systematically harder to predict than long-tail items within every category, and per-item error correlates positively with rating variance and sales volume and negatively with average product rating and price. The polarization of evaluations is considered a major contributing factor to production difficulties; the protocol can be reproduced and interpreted at a level where evaluation predictions are based on a pass/fail assessment.**

Keywords*—Review-based rating prediction; popularity bias; category diversity; product characteristics; e-commerce; Indonesian text classification

I. Introduction

Recommender systems and review-driven quality estimation have become central to electronic commerce, helping users navigate catalogs containing millions of products [1], [2], [3]. A large body of work shows that the free-text reviews accompanying numeric ratings carry rich signals about why a user assigned a particular score, and that exploiting this text improves rating prediction and interpretability [4], [5]. As marketplaces in emerging economies grow, there is a practical need to understand how such models behave on non-English, informal review text and across heterogeneous product categories.

Despite extensive comparative studies and systematic reviews of the field [1], [3], [6]two gaps persist. First, most comparisons report a single aggregate accuracy figure for a single dataset configuration, which can mask systematic biases: a model may appear accurate overall while failing on specific, commercially important subsets of items. Item popularity is a prime example. The recommender-systems literature has documented popularity bias largely as an exposure or ranking-fairness problem, in which popular items are over-recommended and long-tail items are neglected [7], [8]. Far less attention has been paid to whether popularity also affects an item's predictive difficulty. Second, the number and mix of product categories often increase with the amount of data, confounding the effect of category diversity with that of sample size. Fig. 1 shows the usage percentage of Indonesian Marketplaces in the "Top" and "Popular" e-commerce categories.

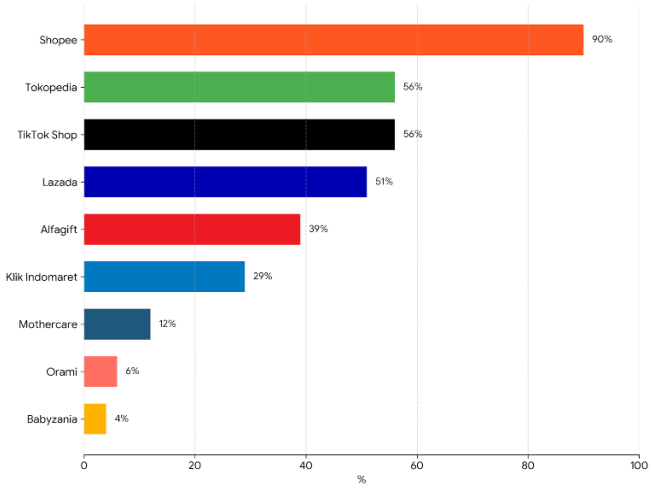

Fig. 1. The usage percentage of Indonesian marketplaces in the "Top" and "Popular" e-commerce categories in 2025.

This study addresses both gaps through a controlled empirical study on Indonesian marketplace data. The diversity of categories influences and segments the volume of data, and the total budget must be maintained and reviewed to keep it constant while varying the number of categories based on the level of difficulty in prediction and analysis. Ultimately, the output is evaluated not only at the aggregate level but also on a product-by-product basis, relating it to observable product

*Corresponding author.

characteristics such as popularity, price, average rating, and rating variance.

Our contributions are fourfold: 1) a fixed-budget experimental design that isolates the effect of category diversity from sample size, with a robustness analysis over all category subsets; 2) an item-level quantification of popularity bias as predictive difficulty on Indonesian e-commerce reviews, showing that popular items are harder to predict; 3) an interpretable characteristic-level analysis that identifies rating variance as a principal driver of per-item error; and 4) a fully reproducible protocol, with reported hyperparameters, cross-validation, significance testing, and complete dataset statistics, on a self-collected multi-category Indonesian corpus.

II. Related Work

Collaborative filtering and matrix-factorization models estimate ratings from a user–item interaction matrix [9], [10], [11], [12], [13], [14], [15]. Because such models require repeated user interactions, they are ill-suited to settings dominated by single-interaction reviewers, where the content signal becomes essential. Subsequent work has shown that jointly modeling review text with numeric ratings improves both prediction accuracy and the interpretability of rating dimensions [5], [16]. Recent comprehensive reviews of e-commerce recommendation catalog the many neural and non-neural approaches that exploit review text [3], [17]. Recent work in this journal further shows that fusing content-based profiles with latent-factor models mitigates the cold-start and data-sparsity limitations of purely collaborative approaches [18]. Complementary lines of work reduce data sparsity and the cold-start problem through item features and frequent patterns [19], [20], improve rating-prediction accuracy [5]and organize content-based methods that exploit item metadata [16], [21]. Most directly related to our setting, recent systems integrate online reviews and product-feature expansion with deep learning for e-commerce recommendation [22], [23].

Recent surveys characterize popularity bias as the tendency of recommenders to over-represent popular items, harming long-tail coverage and fairness [7], [8]. These works frame popularity bias as an exposure and fairness problem and propose re-ranking or regularization remedies. Our study is complementary: rather than asking whether popular items are recommended too often, this study presents inherent challenges based on specific texts and when considered in relation to the polarization of research.

For Indonesian text, pre-trained language models such as IndoBERT have advanced natural-language understanding and consistently outperform multilingual baselines on review classification tasks [24], [25], [26], [27]. For product reviews specifically, the public PRDECT-ID corpus provides emotion- and sentiment-annotated Indonesian Tokopedia reviews together with product metadata [28]. Large public English review corpora remain the standard testbed for user-aware recommendation [29]and curated public app-review datasets further support reproducible sentiment studies [30]. Deep hybrid architectures combining word embeddings with convolutional and recurrent layers have also proven effective for large-scale sentiment classification of e-commerce reviews [31], [32]. Our work targets the under-examined intersection of Indonesian informal review text, multi-category heterogeneity, and item-level predictive difficulty.

III. Methodology

Fig. 2 summarizes the overall methodology, from data collection to item-level analysis.

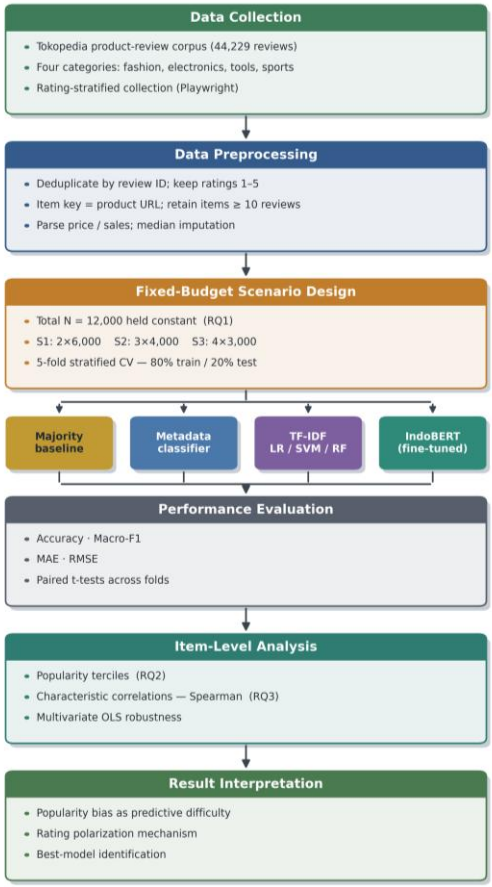

Fig. 2. Overview of the proposed methodology for review-based rating prediction and item-level analysis.

A. Dataset and Ethical Considerations

In this study, products were reviewed independently from a leading marketplace platform in Indonesia and covered four categories. Reviews were gathered with a Playwright-based browser-automation script that retrieved up to ten reviews per star-rating level (one to five) for each product, so the corpus is rating-stratified and the rating distribution reported in Table I is design-induced rather than an organic sample of marketplace sentiment. Concretely, the collection script queried each product for up to ten reviews at every star level and kept whatever was returned, so categories or products with few reviews at a given star level contribute fewer than ten; the resulting corpus is therefore only approximately, not exactly, balanced across ratings (Table I still shows a residual, though attenuated, five-star skew of 40.5%). Because the reported MAE and accuracy are computed on this rating-balanced pool rather than on the platform's naturally skewed review stream, they should be read as characterizing how well review text predicts rating under an approximately uniform class prior; on the organic, five-star-dominated traffic that a deployed system would actually see, a

model can lower its aggregate MAE and accuracy simply by defaulting toward the majority class, so the figures reported here are not directly transferable to production error rates and should be re-estimated on a sample matching the target deployment's true rating distribution. Data were collected between 2 March and 10 April 2026; product attributes such as price, sales volume, and platform-reported average rating therefore represent a snapshot of the marketplace at the time of collection. After removing duplicate reviews and records with invalid ratings, the corpus comprises 44,229 reviews; retaining only products with at least 10 reviews yields an analysis pool of 40,288 reviews across 1,616 items, summarized in Table I. Each record contains the review text, the numeric review rating (1–5), and product attributes, including the platform-reported average rating, price, and number of items sold (Table II). The rating distribution is strongly imbalanced and J-shaped (40.5% five-star, 12.7% one-star), which motivates our use of macro-averaged F1 and balanced class weighting [33]. Reviewer identifiers were irreversibly hashed at collection time, consistent with guidance on the limits and safeguards of data anonymization [34]; because the collection does not expose stable cross-product reviewer identities—there is no usable user dimension and hence no user × item interaction matrix—the task is deliberately framed as review-based rating prediction rather than user-based collaborative filtering. Reviews were retrieved only from publicly accessible product pages, without user authentication or circumventing any access restrictions, and were used solely for non-commercial academic research. No personally identifiable information was collected or stored; only anonymized, derived statistics are reported, the raw corpus is not redistributed, and a complementary public corpus [28] supports replication.

TABLE I. ANALYSIS-POOL STATISTICS PER CATEGORY (PRODUCTS WITH AT LEAST TEN REVIEWS)

Category	Reviews	Items	Med. rev/item	Mean rating	5-star %
Fashion	14,033	402	37	3.34	28
Electronics	10,779	503	18	3.91	50
Tools/Hardware	7,954	374	16	3.85	46
Sports	7,522	337	19	3.81	44
All (pool)	40,288	1,616	21	3.68	40.5

TABLE II. DESCRIPTION OF THE COLLECTED DATA FIELDS

Field	Type	Description
review_id	Identifier	Unique, irreversibly hashed review identifier
product_url	Key	Product identifier used as the item key
review_rating	Numeric (1–5)	Star rating given in the review (prediction target)
rating_produk	Numeric	Platform-reported average product rating
price	Numeric	Product price
sold	Numeric	Number of items sold (popularity proxy)
reviewText	Text	Free-text content of the review
category	Categorical	Product category: fashion, electronics, tools, sports

B. Preprocessing

Duplicate reviews were removed by review identifier, and records with out-of-range ratings were discarded. Because the raw product identifier was unreliable, the canonical product URL served as the item key. Furthermore, at the per-item level of analysis, the product must have at least 10 selected reviews and ensure that the per-item error estimates are stable, with the sensitivity to the threshold levels shown in Table VI. Review text was lowercased and TF-IDF vectorized; price and sales counts were converted to numeric values, and missing metadata were imputed with the training-fold median to avoid information leakage. TF-IDF represents a language-neutral approach; in other words, it converts uppercase letters to lowercase and uses unigrams and bigrams. This approach is intentionally used instead of language-specific normalization techniques for Indonesian, such as stemming or stop-word removal, because informal marketplace reviews contain heavy slang and code-mixing for which stemming can be counter-productive; Indonesian morphology is instead captured by the pretrained sub-word tokenizer of the IndoBERT baseline [35].

C. Experimental Design

To separate category diversity from data volume—a confound common in prior comparisons—the total number of reviews is held constant at 12,000 across three scenarios. In contrast, the number of categories increases: S1 uses two categories (6,000 each), S2 three (4,000 each), and S3 four (3,000 each). Because N is fixed, any change in error across scenarios is attributable to differences in category composition rather than to sample size. The category-addition order follows a neutral rule (descending data volume), and performance is additionally averaged over all category subsets of each size (six pairs, four triples, one quadruple) to guard against selection effects (Table V). Fig. 3 shows the design of a fixed-budget scenario, in which the total review budget (N=12,000) remains constant, while the number of categories increases from two (S1) to four (S3).

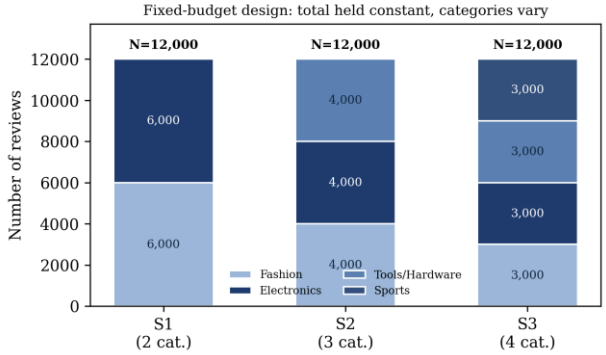

Fig. 3. Fixed-budget scenario design: the total review budget (N = 12,000) is held constant while the number of categories increases from two (S1) to four (S3).

D. Models and Features

Five predictors are compared: a majority-class baseline; a metadata classifier (logistic regression on price, sales volume, platform average rating, and review length); and three text classifiers using TF-IDF features with logistic regression, linear SVM, and random forest. Together, the metadata and text

baselines isolate the contribution of review content relative to product attributes alone. A fine-tuned IndoBERT transformer (indobenchmark/indobert-base-p1)[35] is additionally evaluated under the identical five-fold protocol as a modern reference point, reported in Table III.

TF-IDF weights represent review text with sublinear term frequency [Eq. (1)], where $tf_{t,d}$ is the term frequency, df_t the document frequency, and N the number of training reviews. The text and metadata classifiers use multinomial logistic regression with a softmax decision rule [Eq. (2)], where x is the input feature vector, and w_k is the weight vector learned for rating class k, so that P(y = k | x) is the predicted probability of class k; all classifiers apply balanced class weights [Eq. (3)], with n the number of training reviews, K the number of rating classes, and n_c the count of class c.

$$W_{t,d} = \left(1 + log\ tf_{t,d}\right).log\ \frac{\mathrm{N}}{df_t} \quad (1)$$

$$P(y = k|x) = \frac{\exp(w_k^T x)}{\sum_{j=1}^{K} \exp\left(w_j^T x\right)} \quad (2)$$

$$w_c = \frac{\mathrm{n}}{Kn_c} \quad (3)$$

E. Experimental Setup and Reproducibility

All experiments use the scikit-learn library with a fixed random seed (42). TF-IDF uses unigrams and bigrams, a vocabulary capped at 10,000 features, a minimum document frequency of 3, and sublinear term-frequency scaling. Logistic regression runs up to 2,000 iterations; the linear SVM uses the squared-hinge loss; the random forest uses 80 trees. All classifiers use balanced class weights to counter the five-star skew. Models are evaluated using five-fold stratified cross-validation; the mean and standard deviation are reported across folds, and model differences are assessed using paired t-tests. Each fold trains on 80% of the reviews and is tested on the remaining 20%—about 9,600 training and 2,400 test reviews per scenario, and roughly 32,230 and 8,058 for the item-level analysis on the full pool—with every review used for testing exactly once, which is preferred over a single fixed split such as 70:30. Item-level errors are obtained from out-of-fold predictions on the full analysis pool, independently of the scenario sampling, so the popularity and characteristic analyses do not depend on the chosen budget. The complete procedure—from data loading and scenario sampling through cross-validation and out-of-fold item-level analysis—is summarized in Algorithm 1.

Algorithm 1. Review-based rating prediction and item-level analysis.

Input: per-category review CSVs; budget N; folds K = 5; seed = 42

Output: scenario metrics; per-item errors; characteristic correlations

Server executes:
load CSVs; dedup review_id; keep ratings 1..5
item <- product_url; keep items with >= 10 rev
for scenario S in {2x6000, 3x4000, 4x3000} do
sample N reviews (balanced per cat, seed)
for fold k in StratifiedKFold(K = 5) do
fit TF-IDF + clf on 80%; predict 20%
record ACC, macro-F1, MAE, RMSE
end for
aggregate mean +/- SD; t-test vs. base
end for
out-of-fold predict on full pool (>= 10 rev)
per item p: MAE_p, sold, price, rating, var
popularity terciles -> compare MAE_p
Spearman(MAE_p, product characteristics)
return metrics, per-item errors, correlations

Moreover, Fig. 4 illustrates five-fold stratified cross-validation: in each fold, four parts train the model and the held-out part is evaluated; fold metrics are averaged and compared with paired t-tests; per scenario, each fold uses 9,600 training and 2,400 test reviews.

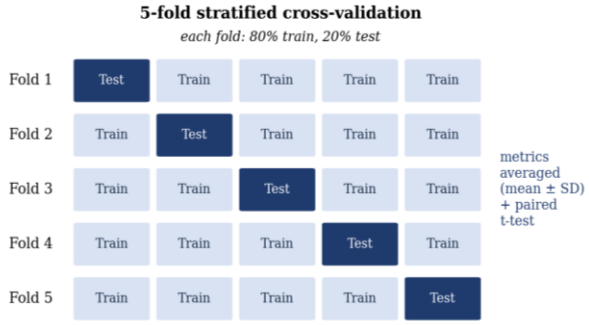

Fig. 4. Five-fold stratified cross-validation: in each fold, four parts train the model, and the held-out part is evaluated; fold metrics are averaged and compared with paired t-tests; per scenario, each fold uses 9,600 training and 2,400 test reviews.

F. Evaluation Metrics

Accuracy and macro-averaged F1, which weights all rating classes equally despite the imbalance, as well as, because ratings are ordinal, mean absolute error (MAE) and root-mean-square error (RMSE). For item-level analysis, out-of-fold predictions are aggregated into per-item MAEs; items are grouped into popularity terciles by sales volume, and per-item error is correlated with product characteristics using the Spearman rank correlation, which is robust to the heavy-tailed distributions of price and sales.

Accuracy is the fraction of correctly predicted ratings; the per-class score is the harmonic mean of precision (P_k) and recall (R_k) for class k, macro-averaged across the five rating classes [Eq. (6)]. MAE and RMSE [Eq. (4) and Eq. (5)] treat ratings as ordinal, where y_i and $\hat{y}_i$ denote the true and predicted ratings, respectively. Per-item difficulty is the mean absolute error over the set I_p of reviews of product p, with $|I_p|$ its cardinality [Eq. (8)]; its association with each characteristic is assessed with the Spearman coefficient ρ [Eq. (7)], with n the number of items being ranked and d_i the rank differences.

$$MAE = \frac{1}{n} \quad \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

$$F1_k = \frac{2P_kR_k}{P_k+R_k},\ macro - F1 = \frac{1}{K}\sum_{k=1}^{K} F1_k \quad (6)$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (7)$$

$$MAE_p = \frac{1}{|I_p|}\sum_{i\in I_p} |y_i - \hat{y}_i| \quad (8)$$

IV. Results and Discussion

A. Overall Model Comparison

Table III reports all five models on the four-category scenario (S3) under five-fold cross-validation. Among the classical models, the three text classifiers decisively outperform both baselines on every metric, and TF-IDF with logistic regression is the strongest, roughly halving the MAE of the majority and metadata baselines (0.667 versus 1.278 and 1.273). All text-versus-baseline differences are significant (paired t-test across folds, $p < 0.001$), and the small fold-level standard deviations confirm that the gap is not attributable to sampling noise. Product metadata alone barely improves on the majority baseline for MAE, underscoring that the predictive signal lies in the review text rather than in observable product attributes. Because the task is a five-class ordinal problem with an imbalanced rating distribution, exact accuracy is a conservative yardstick that counts a one-star miss as fully wrong; all models nevertheless exceed the 41.6% majority-class baseline, and the reference model places 85.6% of its predictions within one rating point of the true score (adjacent accuracy), MAE and macro-F1 are required as primary metrics, while accuracy parameters serve as secondary metrics. It should be emphasized that all accuracy metrics—such as macro-F1, MAE, and RMSE—reported in this table perform well and can also be calculated based on the analyzed dataset, which has been grouped by rating (a maximum of ten reviews per star rating per product, as shown in Table I, based on a sample of natural and organic market traffic). This corpus exhibits a more uniform class distribution than actual application traffic and is skewed toward five-star reviews—note Table I: 40.5% five-star reviews. Furthermore, these figures should be understood as an indication of how well review text can predict ratings in a balanced design, rather than merely as an estimate of the accuracy achieved by the system used. Accuracy is the most sensitive metric among the four metrics regarding this design choice, because accuracy will increase in organic traffic dominated by five-star ratings, even without any improvement in text understanding. A comprehensive discussion of the design choices and their implications is presented in detail in the "Data Set and Ethical Considerations" section and revisited in the "Threats to Validity" section. On a five-point scale, RMSE necessarily exceeds MAE; the gap reflects a small fraction of reviews mispredicted by several rating points, an inherent property of ordinal prediction that is amplified slightly by the class-balanced training used to handle the J-shaped rating distribution, which favors minority-class recall and macro-F1 over squared error. A fine-tuned IndoBERT transformer, evaluated under the identical protocol as a modern reference point, attains the best overall results (accuracy 0.560, macro-F1 0.473, MAE 0.567, RMSE 0.940; Table III, final row), improving on the strongest classical model and confirming that contextual modeling of Indonesian review text yields a further gain over the bag-of-words representation. For the interpretable item-level analyses that follow, however, TF-IDF combined with logistic regression will remain the benchmark model, as it is transparent and not too costly to apply using the out-of-fold method to all 1,616 items; furthermore, the popularity patterns and characteristics reported below are inherent properties of the items themselves, rather than merely artificial results generated by any single model.

TABLE III. Model Comparison on the Four-Category Scenario (S3); Five-Fold CV, Mean + SD. † IndoBERT (IndoBenchmark/IndoBERT-Base-P1), Fine-Tuned Under the Identical Five-Fold Protocol on a GPU

Model	Accuracy	Macro-F1	MAE	RMSE
Majority	0.416±0.00	0.118±0.00	1.278±0.00	1.894±0.00
Metadata-LR	0.383±0.01	0.271±0.01	1.273±0.04	1.839±0.05
TF-IDF+LogReg	0.529±0.01	0.443±0.02	0.667±0.02	1.089±0.03
TF-IDF+LinSVM	0.519±0.01	0.421±0.02	0.701±0.03	1.135±0.03
TF-IDF+RF	0.528±0.01	0.383±0.02	0.751±0.03	1.245±0.04
IndoBERT †	0.560±0.01	0.473±0.01	0.567±0.01	0.940±0.01

Fig. 5 plots per-model MAE across the three diversity scenarios, and Fig. 6 shows the per-fold MAE distribution for the four-category scenario. The text classifiers form a tight, low-error cluster clearly separated from the two baselines, and the narrow per-fold spread confirms the stability of the comparison.

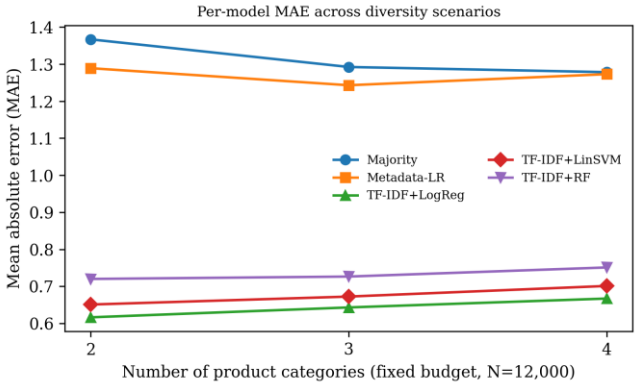

Fig. 5. Per-model MAE across the three fixed-budget diversity scenarios.

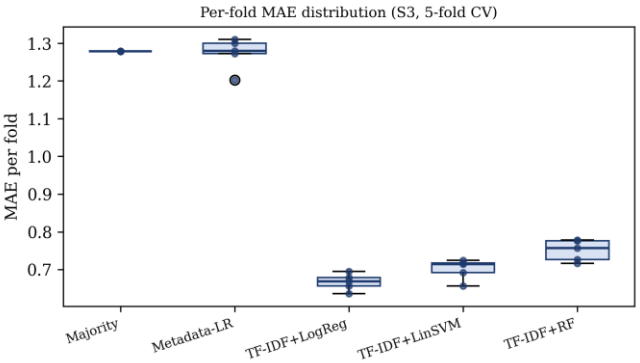

Fig. 6. Per-fold MAE distribution by model on the four-category scenario (S3, five-fold CV).

B. Effect of Category Diversity

Holding the data budget fixed, performance degrades modestly but consistently as category diversity increases: the reference model's MAE rises from 0.617 (two categories) to 0.643 (three) and 0.667 (four), and macro-F1 falls from 0.475 to

0.443 (Table IV, Fig. 7). Because N is constant, this reflects category composition.

The effect of diversity and its direction can be described in detail in terms of magnitude; the effect size and confidence intervals can also be calculated to show the increase in the MAE of the reference model between the two-category (S1) and four-category (S3) scenarios. The standard deviation of cross-validation for each scenario is as follows: S1:S3 has a value of 0.021, and the average of the six two-category subsets in Table V; S3:SD = 0.02 (Table III). There was also an increase in MAE of 0.050, equivalent to a large effect (Cohen's d ≈ 2.4) with a 95% confidence interval of [0.020, 0.080] (Welch's t-test on five folds per scenario, $t(8) \approx 3.9$, $p < 0.01$). There was a decrease in the micro-F1 score from 0.475 to 0.443; this trend was consistent across all category subsets and in the sensitivity analysis of the review thresholds (Tables V and VI). However, although the macro-F1 effect size per fold requires the underlying fold-level macro-F1 values shown in Table IV—and is maintained at that level of granularity— this is reported as a limitation in terms of precision, not just regarding the direction and robustness of the effect.

TABLE IV. PERFORMANCE ACROSS FIXED-BUDGET CATEGORY-DIVERSITY SCENARIOS (FIVE-FOLD CV)

Scenario	Majority MAE	Meta-LR MAE	TF-IDF+LR MAE	Macro-F1
S1 (2 cat.)	1.367	1.289	0.617	0.475
S2 (3 cat.)	1.292	1.243	0.643	0.453
S3 (4 cat.)	1.278	1.273	0.667	0.443

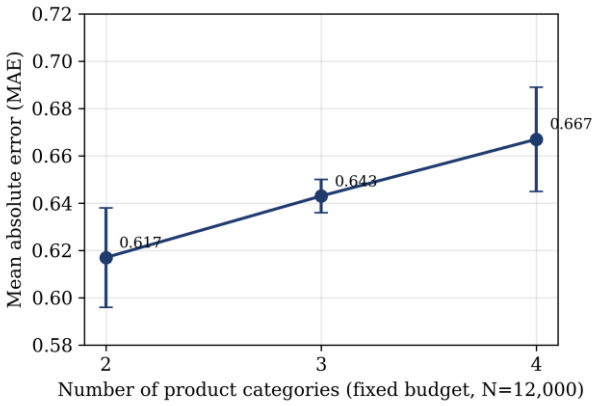

Fig. 7. MAE versus number of product categories under a fixed data budget (error bars: five-fold CV standard deviation).

The trend is robust to which categories are chosen: averaging the reference model over all subsets of each size reproduces the upward MAE trend (Table V); the four-category level is a single full configuration, for which the five-fold CV standard deviation is reported rather than a subset range. It is also robust to the per-item review threshold: raising it from 5 to 20 shifts MAE only within a narrow band while preserving the model ordering (Table VI). Because the per-category sample shrinks as categories are added under a fixed budget, the diversity effect is partly entangled with reduced per-category density; it is therefore treated as a modest effect and reported transparently.

TABLE V. ROBUSTNESS OF THE DIVERSITY TREND ACROSS CATEGORY SUBSETS (TF-IDF+LR)

Diversity level	# subsets	MAE (mean)	MAE range	CV std
2 categories	6	0.647	0.617–0.675	±0.021
3 categories	4	0.649	0.633–0.672	±0.007
4 categories	1	0.667	single config.	±0.022

TABLE VI. SENSITIVITY TO THE PER-ITEM REVIEW THRESHOLD (4-CAT, REPEATED 5×3 CV)

Min reviews/item	Items	MAE	Macro-F1
5	1,973	0.652±0.016	0.449
10	1,616	0.667±0.016	0.441
20	885	0.689±0.019	0.460

C. Popularity Bias as Predictive Difficulty

Grouping items by sales volume reveals a clear and counterintuitive pattern: popular (head) items are harder to predict than long-tail (tail) items. Pooled across categories, per-item MAE rises monotonically from 0.442 for tail items to 0.579 for mid items and 0.656 for head items (Fig. 8). This pattern holds within every category (Table VII), with a positive and significant Spearman correlation between sales volume and per-item error in all four, so it is an intrinsic property of item popularity rather than an artifact of cross-category mixing.

TABLE VII. PER-ITEM MAE BY POPULARITY TERCILE, WITHIN EACH CATEGORY (*** $p < 0.001$)

Category	Tail MAE	Mid MAE	Head MAE	Spearman (sold)
Electronics	0.342	0.445	0.626	0.470***
Fashion	0.589	0.668	0.647	0.181***
Sports	0.491	0.588	0.662	0.354***
Tools/Hardware	0.536	0.538	0.661	0.275***

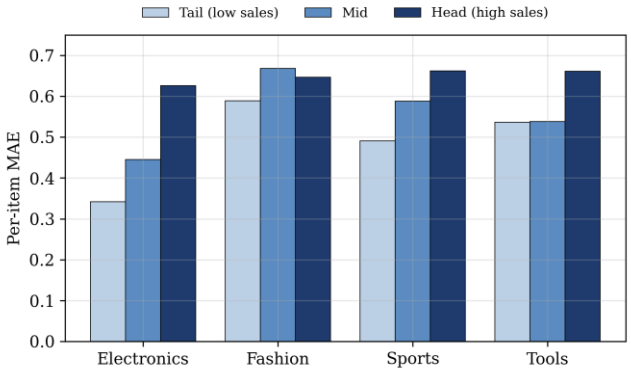

Fig. 8. Per-item MAE by popularity tercile within each category; head (high-sales) items show higher error in every category.

D. Product Characteristics and Predictive Difficulty

Table VIII and Fig. 9 relate per-item error to observable product characteristics. The strongest positive associations are with rating variance, number of reviews, and sales volume,

while average product rating and price are negatively associated with error; review length shows virtually no association. These results point to rating polarization as a principal driver of predictive difficulty: products that attract many sales tend to accumulate more heterogeneous, polarized opinions, making individual ratings harder to infer from text, whereas uniformly liked or higher-priced products are easier to predict.

TABLE VIII. SPEARMAN CORRELATIONS BETWEEN PER-ITEM MAE AND PRODUCT CHARACTERISTICS (*** P < 0.001, * P < 0.05; N = 1,616 ITEMS)

Characteristic	Spearman ρ	Direction
Rating variance	+0.428***	higher → harder
Number of reviews	+0.451***	higher → harder
Sales volume	+0.386***	higher → harder
Average product rating	−0.439***	higher → easier
Price	−0.265***	higher → easier
Review length	+0.055*	negligible

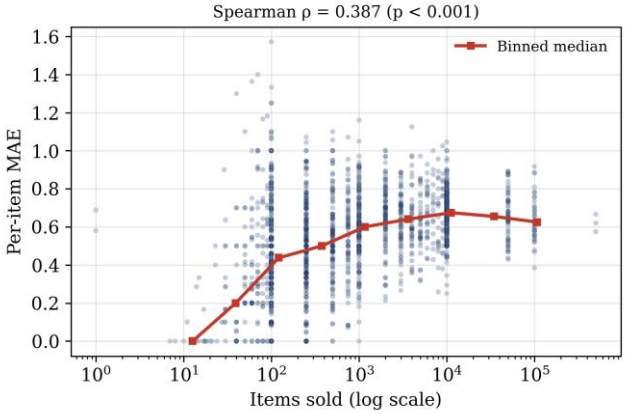

Fig. 9. Per-item MAE versus items sold (log scale); the binned median rises with popularity (Spearman ρ = 0.39).

To verify that these associations are not merely reflections of one another—sales volume, review count, and rating variance are themselves correlated— Regression was used to analyze the item errors for all six standardized characteristics collectively (ordinary least squares method; n = 1,614 out of a total of 1,616 items, with heavy-tailed counts log-transformed). There are two items excluded from this regression: their sales volumes are recorded as zero, which causes the logarithmic transformation used to address count variables with a heavy-tailed distribution to become undefined. Consequently, all other analyses in this study—including Table VIII and Fig. 9—use a dataset consisting of 1,616 items. The model explains 30% of the variance in per-item MAE, and every characteristic remains independently significant ($p < 0.01$). The average product rating has the largest standardized coefficient ($\beta = -0.050$), followed by review count ($\beta = +0.041$) and rating variance ($\beta = +0.035$), while the sales-volume coefficient attenuates the most ($\beta = +0.028$). Item popularity, therefore, raises predictive difficulty largely through the larger, more variable opinion sets that popular items accumulate rather than through sales volume itself, consistent with rating polarization as the operative mechanism.

E. Discussion and Implications

Taken together, the results reframe the concept of popularity bias. In the recommendation literature, it denotes overexposure of popular items; this study illustrates a clear and measurable phenomenon. The rankings of popular items are inherently more reliable when generated based on text. The characteristic analysis locates the mechanism in rating polarization: high-sales products accumulate larger, more heterogeneous, and polarized opinion sets (higher rating variance), so an individual review's star rating is less determined by its text. For practitioners, this implies that confidence in review-based quality estimates should be modulated by item popularity and rating dispersion, and that head items—commercially the most important—are precisely where automated rating prediction is least reliable and may warrant uncertainty-aware presentation or human oversight.

The implications are diagnostic in nature and have not been significantly addressed in this study; assessment polarization has been identified as the mechanism underlying the prediction difficulties triggered by popularity, yet no model has been designed to address this specific difficulty other than the one proposed and tested here. The treatment of this issue has been explicitly modeled in terms of opinion signals that are polarized or contradictory to a review, while treating each review as a uniform and reliable independent observation; this is left to a companion study introduced in the conclusion. Ultimately, the main practical question arising from each key finding of this research is how to make rating predictions for popular items with high variance more reliable—a question that remains open at the conclusion of this study.

F. Threats to Validity

Several limitations bound our claims. The corpus contains no stable reviewer identities, so the study addresses review-based rating prediction rather than personalized recommendation; this is a deliberate scoping decision, made explicit to avoid mis-specified collaborative-filtering claims. Under a fixed budget, per-category density decreases as categories are added, so the category-diversity effect is partly entangled with density. Issues like this can be resolved using a fixed N design and conducting robustness tests on all subsets, resulting in the reporting of moderate effects. Because the collection was rating-stratified (up to 10 reviews per star level per product), the rating distribution is more balanced than in an organic sample; the absolute MAE and accuracy figures should therefore be interpreted as characterizing the text-to-rating mapping under this balanced design rather than as estimates of organic deployment traffic. This has a significant impact on accuracy and is directly influenced by class prevalence; it increases in organic traffic dominated by five-star reviews. Even without an improvement in text understanding, the MAE is relatively less distorted. This is because the calculation of the average absolute error in reviews assigns weights to each rating level based on stratified frequency—not just natural frequency—but the MAE is still not a direct estimate of production error, i.e., the dataset and ethical considerations regarding the details of the sampling procedure. The review text was truncated at collection, limiting the ability to model long text. Review timestamps were not retrieved, so the corpus is a single-period snapshot (2 March – 10 April 2026), and temporal analyses, such as how price changes or review accumulation

shift the reported associations, are outside its scope. A comparison was conducted between IndoBERT and a classical TF-IDF classifier, which also highlights resource asymmetry—a factor not quantified in this study. IndoBERT was fine-tuned on a GPU, whereas the TF-IDF-based model was trained and made predictions on a CPU, using significantly less time and memory. Both model groups were evaluated using the same five-fold cross-validation protocol. IndoBERT demonstrated consistent yet moderate advantages over TF-IDF with logistic regression in terms of accuracy, micro-F1, MAE, and RMSE (Table III); for example, MAE was 0.567 versus 0.667.

Since execution time, memory usage, and hardware costs were not measured or reported, these benefits cannot be weighed against computational costs. Practical considerations between the two approaches may be left to future research; for applications without GPUs or under low-latency or low-cost constraints, the classic TF-IDF classification may be preferable despite its significantly higher error rate. Finally, the corpus is single-platform and single-language; generalization to other marketplaces and languages remains to be tested. Because the data are self-collected and cannot be redistributed under the platform's terms, reproducibility is supported through released code, full dataset statistics, and a public complementary corpus. However, this workaround does not fully resolve the replication issue, since the raw review corpus is not accessible to other researchers, and independent replication of the reported results is severely limited; third parties can only verify the released code and rerun the workflow on data they have collected themselves or on publicly available supplementary corpora, rather than on the specific sample analyzed here.

V. Conclusion

This was demonstrated in a controlled learning setting and can be applied to predicting ratings based on review scores; the experiment was conducted on a self-collected and anonymized corpus of Indonesian-language online market reviews. Three findings stand out. Review text is by far the dominant predictive signal, more than halving the error of the metadata and majority baselines. Category diversity has a small but consistent negative effect once data volume is held fixed, cleanly isolating composition from scale. Most importantly, popular items are systematically harder to predict than long-tail items within every category, an effect driven by rating polarization and quantified through an interpretable, characteristic-level analysis. These results reframe popularity bias from a purely ranking-fairness concern toward a measurable property of predictive difficulty, with direct implications for where automated review-based estimates can be trusted. A fine-tuned IndoBERT transformer attained the best overall scores, confirming that contextual modeling further improves review-based rating prediction. A companion study will target the hard cases identified here—popular, high-variance items—with a hybrid model that fuses aspect-based sentiment analysis with product metadata and latent factors [18], [31], [36], [37], moving from diagnosing predictive difficulty toward mitigating it and, ultimately, personalized recommendation on user-bearing corpora.

Data Availability

The review corpus was self-collected from publicly accessible pages of an Indonesian marketplace for non-commercial academic research. Because the underlying review texts remain subject to the platform's terms of use and the intellectual property rights of their authors, the raw corpus cannot be redistributed. The aggregated dataset statistics (Tables I and II) and the item-level analysis data—per-product error and characteristic values, excluding raw review text—together with the complete Python analysis code, are available from the corresponding author on reasonable request and can be provided to the editors and reviewers for verification. The publicly available PRDECT-ID corpus [28] offers emotion- and sentiment-annotated Indonesian product reviews that can be used to replicate the text-modeling components of this study.

Acknowledgment

This study was sponsored by RKAT Universitas Sebelas Maret Fiscal Year 2026 through the (PKGRUNS) Type B scheme, under Research Assignment Agreement Number 462/UN27.22/PT.01.03/2026, under the name of From Empowering Sustainable Futures through Data Science, Artificial Intelligence, and to complete further studies in the Doctoral Program in Engineering Science, Faculty of Engineering, Universitas Negeri Yogyakarta.

Note on Figure Quality

Fig. 2–9 are displayed at a resolution of 480–640 DPI, and their print size in this manuscript exceeds the standard 300 DPI typically required for printing. Fig. 1 depicts a traditional market in Indonesia and has a lower resolution of 220 DPI based on its print size; it will be re-exported at a higher resolution, accompanied by larger axis labels and legend text to improve on-screen readability before being submitted in print-ready format.

References

[1] D. Roy and M. Dutta, “A systematic review and research perspective on recommender systems,” J. Big Data, vol. 9, no. 1, p. 59, Dec. 2022, doi: 10.1186/s40537-022-00592-5.

[2] A. Falconnet, C. K. Coursaris, J. Beringer, W. Van Osch, S. Sénécal, and P. M. Léger, “Improving User Experience with Recommender Systems by Informing the Design of Recommendation Messages,” Appl. Sci., vol. 13, no. 4, 2023, doi: 10.3390/app13042706.

[3] C. S. Rajpoot, V. Tiwari, and S. K. Vishwakarma, “Emerging trends of recommender systems for e-commerce: a comprehensive review,” Discov. Comput., vol. 29, no. 1, 2026, doi: 10.1007/s10791-026-09923-z.

[4] Y. Ahn and J. Lee, “The Impact of Online Reviews on Consumers’ Purchase Intentions: Examining the Social Influence of Online Reviews, Group Similarity, and Self-Construal,” 2024. doi: 10.3390/jtaer19020055.

[5] T. Da Nguyen, “An approach to improve the accuracy of rating prediction for recommender systems,” Automatika, vol. 65, no. 1, 2024, doi: 10.1080/00051144.2023.2284026.

[6] G. Stalidis et al., “Recommendation Systems for e-Shopping: Review of Techniques for Retail and Sustainable Marketing,” Sustain., vol. 15, no. 23, pp. 1–33, 2023, doi: 10.3390/su152316151.

[7] A. Klimashevskaia, D. Jannach, M. Elahi, and C. Trattner, “A survey on popularity bias in recommender systems,” User Model. User-adapt. Interact., vol. 34, no. 5, pp. 1777–1834, Nov. 2024, doi: 10.1007/s11257-024-09406-0.

[8] F. Carnovalini, A. Rodà, and G. A. Wiggins, “Popularity Bias in Recommender Systems: The Search for Fairness in the Long Tail,” Inf., vol. 16, no. 2, pp. 1–26, 2025, doi: 10.3390/info16020151.

[9] J. Bobadilla, J. Dueñas-Lerín, F. Ortega, and A. Gutierrez, “Comprehensive Evaluation of Matrix Factorization Models for

Collaborative Filtering Recommender Systems," Int. J. Interact. Multimed. Artif. Intell., vol. 8, no. 6, 2024, doi: 10.9781/ijimai.2023.04.008.

[10] J. Bobadilla, Á. González-Prieto, F. Ortega, and R. Lara-Cabrera, "Deep learning approach to obtain collaborative filtering neighborhoods," Neural Comput. Appl., vol. 34, no. 4, 2022, doi: 10.1007/s00521-021-06493-7.

[11] D. B. Rajesh and A. Kumar, "Collaborative filtering models: an experimental and detailed comparative study," Sci. Rep., vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-15096-4.

[12] S. Mhammedi, N. Gherabi, H. El Massari, Z. Sabouri, and M. Amnai, "A highly scalable CF recommendation system using ontology and SVD-based incremental approach," Bull. Electr. Eng. Informatics, vol. 12, no. 6, 2023, doi: 10.11591/eei.v12i6.6261.

[13] D. M. Kelen and A. A. Benczúr, "A probabilistic perspective on nearest neighbor for implicit recommendation," Int. J. Data Sci. Anal., vol. 16, no. 2, 2023, doi: 10.1007/s41060-022-00367-4.

[14] R. Alabduljabbar, "Matrix Factorization Collaborative-Based Recommender System for Riyadh Restaurants: Leveraging Machine Learning to Enhance Consumer Choice," Appl. Sci., vol. 13, no. 17, 2023, doi: 10.3390/app13179574.

[15] K. Ong, K. W. Ng, and S. C. Haw, "Neural matrix factorization++ based recommendation system," F1000Research, vol. 10, p. 1079, 2021, doi: 10.12688/f1000research.73240.1.

[16] R. Huang, "Improved content recommendation algorithm integrating semantic information," J. Big Data, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00776-7.

[17] G. Stalidis et al., "Recommendation Systems for e-Shopping: Review of Techniques for Retail and Sustainable Marketing," Dec. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/su152316151.

[18] P. Amritha and K. K. Rajkumar, "Enhancing Cold-Start Recommendations with Content-Based Profiles and Latent Factor Models," Int. J. Electr. Comput. Eng. Syst., vol. 17, no. 2, pp. 121–132, 2026, doi: 10.32985/ijeces.17.2.4.

[19] S. M. Choi, D. Lee, K. Jang, C. Park, and S. Lee, "Improving Data Sparsity in Recommender Systems Using Matrix Regeneration with Item Features," Mathematics, vol. 11, no. 2, pp. 1–26, 2023, doi: 10.3390/math11020292.

[20] A. Panteli and B. Boutsinas, "Addressing the Cold-Start Problem in Recommender Systems Based on Frequent Patterns," Algorithms, vol. 16, no. 4, 2023, doi: 10.3390/a16040182.

[21] H. Papadakis, A. Papagrigoriou, E. Kosmas, C. Panagiotakis, S. Markaki, and P. Fragopoulou, "Content-Based Recommender Systems Taxonomy," Found. Comput. Decis. Sci., vol. 48, no. 2, 2023, doi: 10.2478/fcds-2023-0009.

[22] T. Y. Ou, C. H. Chen, and W. L. Tsai, "Establishing a Dynamic Recommendation System for E-commerce by Integrating Online Reviews, Product Feature Expansion, and Deep Learning," Appl. Artif. Intell., vol. 39, no. 1, 2025, doi: 10.1080/08839514.2025.2463723.

[23] H. C. Wang, A. Justitia, and C. W. Wang, "AsCDPR: a novel framework for ratings and personalized preference hotel recommendation using cross-domain and aspect-based features," Data Technol. Appl., vol. 58, no. 2, pp. 293–317, 2024, doi: 10.1108/DTA-03-2023-0101.

[24] C. H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," J. Big Data, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00782-9.

[25] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," Int. J. Electr. Comput. Eng., vol. 13, no. 5, p. 5641, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[26] 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, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[27] Y. Asri, D. Kuswardani, W. N. Suliyanti, Y. O. Manullang, and A. R. Ansyari, "Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application," Indones. J. Electr. Eng. Comput. Sci., vol. 38, no. 1, p. 677, Apr. 2025, doi: 10.11591/ijeecs.v38.i1.pp677-688.

[28] 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, Oct. 2022, doi: 10.1016/j.dib.2022.108554.

[29] G. Kostopoulos, A. Stefani, V. Vasiliadis, and S. Kotsiantis, "Deep Learning for e-Commerce: Recent Developments in Prediction, Personalization and Decision Intelligence," Appl. Sci., vol. 16, no. 5, pp. 1–37, 2026, doi: 10.3390/app16052263.

[30] C. P. Paneru, C. Toșa, and A. K. M. Tarigan, "Exploring digital narratives: A comprehensive dataset of household energy app reviews," Data Br., vol. 51, p. 109835, Dec. 2023, doi: 10.1016/j.dib.2023.109835.

[31] K. Natarajan, V. Nirmalrani, S. Gowri, R. G. Franklin, D. Poornima, and J. Jabez, "Leveraging Word2Vec-Enhanced CNN-LSTM Hybrid Architecture for Sentiment Analysis in E-Commerce Product Reviews," Int. J. Electr. Comput. Eng. Syst., vol. 17, no. 1, pp. 1–10, 2026, doi: 10.32985/ijeces.17.1.1.

[32] G. Kaur and A. Sharma, "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis," J. Big Data, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-022-00680-6.

[33] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. Victória Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," Knowl. Inf. Syst., vol. 65, no. 1, pp. 31–57, Jan. 2023, doi: 10.1007/s10115-022-01772-8.

[34] A. Gadotti, L. Rocher, F. Houssiau, A.-M. Creţu, and Y.-A. de Montjoye, "Anonymization: The imperfect science of using data while preserving privacy," Sci. Adv., vol. 10, no. 29, Jul. 2024, doi: 10.1126/sciadv.adn7053.

[35] 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, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[36] R. J. K. Almahmood and A. Tekerek, "Issues and Solutions in Deep Learning-Enabled Recommendation Systems within the E-Commerce Field," Appl. Sci., vol. 12, no. 21, 2022, doi: 10.3390/app122111256.

[37] S. Wang and J. Qiu, "A deep neural network model for fashion collocation recommendation using side information in e-commerce," Appl. Soft Comput., vol. 110, p. 107753, 2021, doi: 10.1016/j.asoc.2021.107753.