Vol. 8, No. 1, February 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Analyzing the Impact of Review Sentiment on Carpentry Product Sales: Evidence from Tokopedia

Agung Chandra Kharisma[1], Muhammad Haykal Alfariz Saputra[2], Ali Ibrahim[3]*, Mira Afrina[4]

[1,2,3,4]Master of Computer Science, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

Received:
December 18, 2025
Revised:
February 4, 2026
Accepted:
February 24, 2026
Published:
March 5, 2026

Corresponding Author:
Author Name*:
Ali Ibrahim
Email*:
aliibrahim@unsri.ac.id

DOI:
10.63158/journalisi.v8i1.1412

Abstract. The rapid growth of e-commerce in Indonesia has increased the importance of consumer reviews as signals influencing purchasing decisions. This study examines the relationship between review sentiment and sales performance in the carpentry tools category on Tokopedia. Using a 2019 Kaggle dataset consisting of 1,826 reviews across approximately 60 products, we apply an NLP-based pipeline to classify review sentiment into positive, neutral, and negative categories. Sentiment labeling combines rating-based rules and a TF-IDF + Logistic Regression baseline, with additional evaluation using IndoBERT. Product-level metrics—including the proportion of positive sentiment (pos_share), average rating, and units_sold (sales proxy)—are analyzed using descriptive statistics, correlation analysis, and cross-sectional OLS regression. The findings reveal that, in this snapshot dataset, the association between positive sentiment share and log(units_sold + 1) is weak and statistically limited, suggesting that sales variation cannot be explained solely by sentiment polarity or average ratings without considering other commercial factors. These results highlight the importance of incorporating contextual variables and temporal design in future research. Practically, the study suggests that sellers should monitor not only sentiment polarity but also the informational richness of reviews to strengthen reputation management strategies.

Keywords: Sentiment analysis, online reviews, e-commerce, utilitarian products, Tokopedia, IndoBERT

1. INTRODUCTION

Indonesia's e-commerce boom has made online marketplaces like Tokopedia increasingly crowded, where many sellers offer similar products at comparable prices and delivery speeds. In that environment, consumer reviews become one of the few highly visible signals that help buyers separate "safe" choices from risky ones. Reviews do more than record satisfaction after the fact—they operate as public evidence of product quality and seller reliability, shaping trust in settings where buyers cannot physically inspect items. For utilitarian products such as carpentry tools, where performance failures can translate into wasted time, damaged materials, or safety issues, the informational role of reviews is especially consequential. Yet the practical question remains: do more positive reviews actually translate into higher sales at the product level, or do they mainly influence perceptions without strongly moving marketplace outcomes?

A substantial body of research shows that review valence is strongly linked to consumer responses such as perceived quality, trust, and purchase intention. Meta-analytic findings consistently position review sentiment among the most influential predictors of buying intention, sometimes outperforming other decision antecedents [1], [2]. Experimental studies further indicate that consumers often rely on the textual substance of reviews—not just star ratings—when forming judgments [3]. At the same time, the literature also warns that mixed or inconsistent reviews can reduce purchase intention, particularly when credibility cues are weak or when buyers perceive manipulation or low diagnosticity [4]. These findings create a broad expectation that positive sentiment should be associated with stronger market performance.

Theoretically, this expectation is supported by multiple frameworks that converge on the idea that reviews shape behavior by reducing uncertainty. Electronic Word of Mouth (e-WOM) explains how consumer-generated content reduces perceived risk and helps establish trust in online transactions [13]. The Stimulus–Organism–Response (S-O-R) model conceptualizes review content as a stimulus that alters internal cognitive and affective states, which then influence responses such as purchase decisions [1]. Social Influence Theory similarly suggests that consumers lean on peer evaluations as heuristics under uncertainty, especially when product evaluation requires technical knowledge or prior experience [4]. Taken together, these perspectives imply that review sentiment

should not only affect intentions in controlled settings but also be reflected in observable marketplace outcomes.

Despite this strong theoretical and behavioral foundation, a critical empirical gap persists in the Indonesian marketplace context. Recent progress in Indonesian NLP—particularly transformer-based approaches such as IndoBERT and benchmark resources such as IndoNLU—has enabled more accurate sentiment classification on Indonesian-language corpora [5], [8]. Studies using Tokopedia-related text data between 2022 and 2025 report strong classification performance with both conventional pipelines (e.g., TF-IDF + Logistic Regression/SVM) and transformer models [6], [9], [10]. However, most work stops at model accuracy, treating sentiment analysis as an end in itself rather than a tool for explaining marketplace behavior. In other words, we increasingly know how to classify sentiment well, but we still know relatively little about whether aggregated sentiment signals—such as the share of positive reviews derived from textual content—are meaningfully associated with sales performance across products in real marketplace conditions.

This limitation is particularly important because many local studies and practical dashboards still rely heavily on average star ratings, which can mask nuance in textual feedback and may be less diagnostic in categories where small performance issues matter. Carpentry tools represent a useful test case: they are utilitarian, performance-driven products evaluated on durability, precision, safety, and reliability. Buyers often seek experiential detail (e.g., "the blade dulls quickly," "the drill overheats," "the handle is uncomfortable") that ratings alone may not convey. If review sentiment truly functions as a reputational and informational signal, then NLP-derived sentiment—grounded in textual meaning rather than numeric labels—should help explain why some products sell substantially more than others.

Accordingly, this study examines the cross-sectional relationship between review sentiment and sales performance for carpentry products on Tokopedia, using units_sold as a sales proxy. Review-level sentiment is estimated using an NLP pipeline (a TF-IDF + Logistic Regression baseline with an IndoBERT comparison), then aggregated into product-level indicators (e.g., proportion of positive sentiment) and evaluated against sales metrics. The core aim is not merely to demonstrate sentiment classification

feasibility, but to test whether sentiment-derived measures carry explanatory power for real marketplace outcomes.

This study makes two contributions. First, it extends Indonesian e-commerce analytics by linking NLP-derived sentiment indicators to a concrete, product-level sales proxy in a utilitarian category where functional performance and risk reduction are central. Second, and more importantly, it offers a critical empirical test of a commonly assumed relationship: that more positive review valence reliably translates into stronger sales. By evaluating whether aggregate sentiment measures materially explain sales variation (rather than assuming they do), this work clarifies whether sentiment polarity alone is a dominant driver of sales performance in Tokopedia's carpentry segment—or whether its influence is more limited once translated from consumer psychology into marketplace reality.

2. METHODS

This study is designed to operationalize review sentiment from Indonesian-language text and then test whether that sentiment—once aggregated at the product level—helps explain cross-sectional differences in sales performance for carpentry tools on Tokopedia. This "text → sentiment → product indicators → sales association" pipeline is deliberately aligned with the Introduction's central gap: many studies in Indonesia demonstrate strong sentiment-classification performance, but fewer examine whether NLP-derived sentiment metrics actually correspond to observable marketplace outcomes. Because Tokopedia reviews frequently include informal expressions, slang, abbreviations, and mixed writing styles, representation choice matters: embedding-based approaches (e.g., Word2Vec, FastText, or transformer embeddings such as IndoBERT) generally capture semantic relationships better than sparse bag-of-words features like TF-IDF, which can be brittle under linguistic variability [5]. At the same time, TF-IDF paired with linear classifiers remains a strong, efficient baseline for short text, making it useful for benchmarking model complexity against practical performance constraints [5].

In addition to overall polarity, reviews of utilitarian products often contain aspect cues (e.g., power, durability, ergonomics, safety), where dissatisfaction may be localized rather than global. Aspect-Based Sentiment Analysis (ABSA) is therefore a natural extension

because it can isolate which functional attributes drive sentiment and would yield more diagnostic guidance for sellers [11]. While ABSA is not the primary modeling approach here, its relevance is acknowledged to frame future extensions consistent with utilitarian evaluation criteria. Model selection in this study explicitly balances accuracy and computational efficiency: Logistic Regression and SVM are stable for high-dimensional text representations, whereas transformer-based IndoBERT can encode richer context at higher computational cost [5]. To ensure that product-level inference is not driven by measurement artifacts from a single modeling choice, classification validity is checked by comparing a baseline pipeline against a transformer pipeline, evaluating whether added complexity produces meaningful gains on this dataset [5].

2.1. Research Design

This research uses a quantitative, cross-sectional design with the product as the final unit of analysis. Sentiment is first predicted at the review level, then aggregated into product-level indicators that can be related to a sales proxy. Because the dataset represents a single snapshot (2019), the analysis is intentionally framed as cross-sectional association rather than a temporal model or causal estimate. This design is still informative for the study's objective: testing whether products with more favorable aggregated sentiment differ systematically in units_sold, which speaks directly to the practical assumption that "more positive reviews → better marketplace performance" in Tokopedia's carpentry segment.

2.2. Data and Sampling

Data were sourced from a publicly available Kaggle dataset containing Tokopedia product reviews (2019 snapshot). The dataset was filtered to the carpentry category to match the Introduction's emphasis on utilitarian, performance-oriented products. After filtering, the analytical sample contains 1,826 reviews across approximately 60 products, including review text, rating (1–5), product_id, and a cumulative sales proxy ("sold"). Observations with incomplete review text or missing sales information were excluded to avoid distorted sentiment measurement or incomplete outcome reporting. Importantly, because the sold field is cumulative and not time-stamped, it is treated as a sales performance proxy rather than a direct measure of contemporaneous transactions; this motivates both the cross-sectional framing and the use of log transformation in later modeling to reduce skewness.

2.3. Research Variables

Table 1 operationalizes how each construct is measured and linked across the review-level and product-level stages. The key conceptual move—central to the Introduction's gap—is that sentiment is derived from textual content and then aggregated into a product-level signal (pos_share) rather than relying only on star ratings. At the same time, rating_mean is retained as a comparator to separate the contribution of text-derived sentiment from the more commonly used numeric rating signal.

Table 1. Research variables and operational definitions

Variable	Operational Definition	Scale	Source/Column
Sentiment Label	Polarity of classified reviews: negative (–1), neutral (0), positive (+1). Obtained through two channels: (A) silver label based on ratings. (B) supervised from text	Nominal	text, rating
pos_share	Proportion of positive reviews per product	Ratio (0–1)	label_sentimen (hasil model)
rating_mean	Average star rating per product	Ordinal (1–5)	Rating
units_sold	Aggregated number of units sold per product as a proxy for sales performance	Ratio	Sold
n_reviews	Number of reviews per product	Ratio	Text

2.4. Sentiment Classification Procedure

Sentiment labeling and prediction were implemented in two stages to create a practical, review-level sentiment signal that can later be aggregated for sales analysis. First, a silver-label mapping was generated from ratings using explicit rules (≥4 positive, 3 neutral, ≤2 negative). This step provides an initial label source where fully manual annotation is unavailable, enabling supervised learning while remaining transparent about label provenance. Second, supervised sentiment classification was performed using TF-IDF (1–2 grams) with Logistic Regression and Support Vector Machine models. IndoBERT was included as a transformer-based comparison model to test whether contextual

embeddings improve sentiment measurement under Indonesian review language variability [5]. The dataset was split using a stratified 80:20 train–test scheme to preserve class proportions, and model performance was evaluated with Accuracy and Macro F1-score, supplemented by a Confusion Matrix to diagnose systematic misclassification patterns (e.g., neutral vs. positive confusion), which is especially relevant when later aggregating polarity into product-level shares.

2.5. Model Evaluation and Sales Association Analysis

To align with the study's core question—whether sentiment explains sales variation—classification evaluation and marketplace modeling are treated as linked but distinct steps. Sentiment classification (review-level) is evaluated via Accuracy, Macro F1, and the Confusion Matrix under the stratified 80/20 scheme. Then, sentiment predictions are aggregated to the product level to compute pos_share, alongside rating_mean, n_reviews, and units_sold. The sales proxy is transformed as log(units_sold + 1) to reduce right-skew and stabilize variance, which is common in marketplace sales distributions. The primary cross-sectional model is estimated using OLS as shown in Equation 1.

$$\log(\text{units\_sold} + 1) = \beta_0 + \beta_1 \cdot \text{pos\_share} + \beta_2 \cdot \text{rating\_mean} + \gamma' \cdot \text{Control} + \varepsilon \quad (1)$$

Controls are drawn from available variables that plausibly correlate with visibility and demand; at minimum, n_reviews is included as a product-level control because review volume can proxy for popularity and exposure. OLS assumptions are evaluated through residual inspection (normality) and heteroscedasticity testing (Breusch–Pagan). Where heteroscedasticity is indicated, robust standard errors (HC1/HC3) are reported, and quantile regression is used as a robustness check in the presence of heavy-tailed outcomes. This modeling choice directly reflects the Introduction's intent to *critically test* whether aggregate sentiment is a dominant explanatory factor, rather than assuming it is.

2.6. Research Flow Diagram

Figure 1 summarizes the end-to-end workflow and clarifies how the unit of analysis shifts from reviews to products. As shown in Figure 1, the process begins with dataset acquisition (Tokopedia 2019 snapshot) and carpentry-category filtering, followed by cleaning and extraction of key fields (review text, rating, product_id, and cumulative units

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

sold). Next, sentiment is estimated at the review level using the two-stage approach: (i) rating-derived silver labels, then (ii) supervised classification using TF-IDF + Logistic Regression/SVM, with IndoBERT as a comparative transformer model [5]. The resulting review-level sentiment predictions are aggregated to construct product-level indicators—most importantly pos_share—alongside rating_mean, n_reviews, and log(units_sold + 1). The final stage applies Pearson correlation and cross-sectional regression to test whether variation in aggregated sentiment is associated with variation in sales performance, with diagnostics and robustness checks to validate inference.

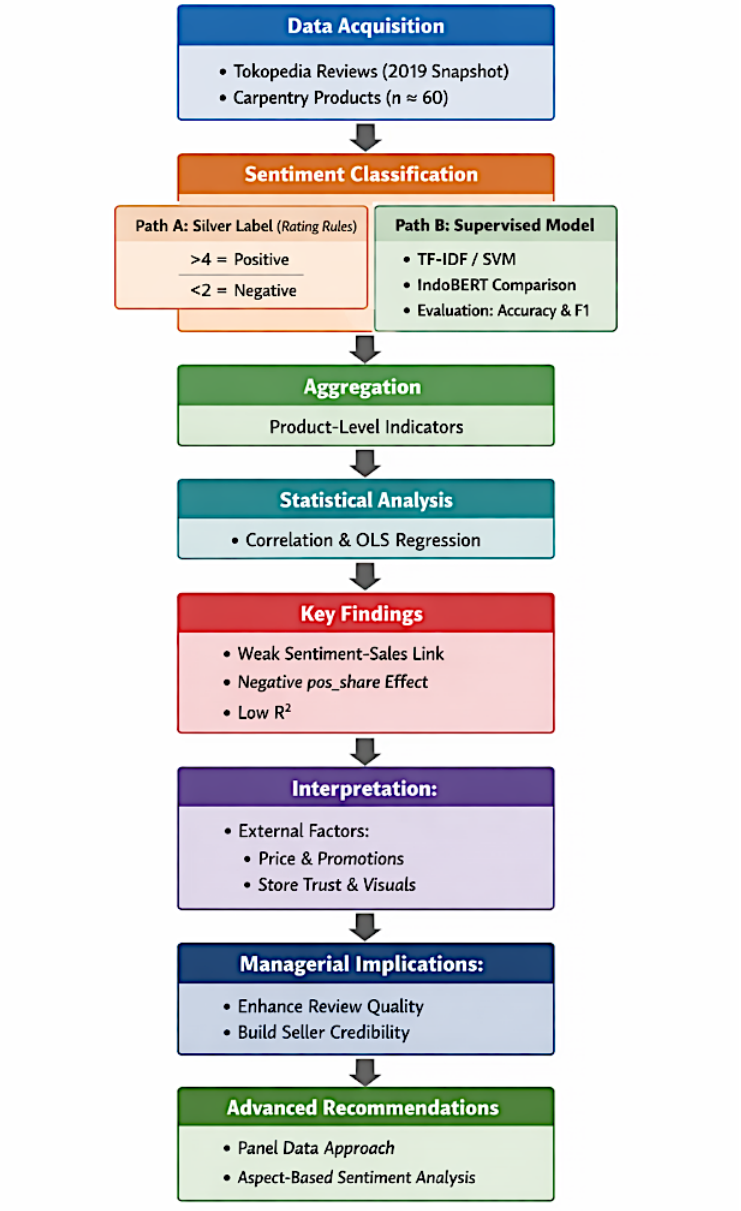

Figure 1. Structured Research Flow

2.7. Limitations and Reproducibility

Consistent with the cross-sectional framing described earlier, the dataset lacks temporal stamps; therefore, dynamic relationships (e.g., sentiment changes preceding sales changes) and causal inference are out of scope. In addition, units_sold is cumulative and may partially reflect listing age, pricing, promotion intensity, search-ranking exposure, or store-level credibility signals. These limitations are explicitly acknowledged so that findings are interpreted as associations rather than causal effects, matching the Introduction's emphasis on empirically testing the strength of the assumed sentiment–performance link under real marketplace constraints.

All experiments were implemented in Python (pandas, scikit-learn, statsmodels) with a fixed random seed (seed = 42) to ensure reproducibility of train–test splits and model results. Preprocessing, labeling, and modeling scripts are provided in an appendix, along with two derivative datasets: (1) the carpentry-category subset and (2) the product-level aggregation file. A non-GPU environment is sufficient for TF-IDF + linear baselines, while a GPU is recommended for IndoBERT fine-tuning to reduce training time, consistent with the computational trade-offs noted earlier [5].

3. RESULTS AND DISCUSSION

This section evaluates whether NLP-derived review sentiment—aggregated as product-level indicators—helps explain cross-sectional variation in sales performance for carpentry products on Tokopedia. While prior literature generally reports that review valence is a strong driver of purchase intention and shopping behavior [1], [2], the empirical contribution of this study is to test that assumption using an observable marketplace proxy (units_sold) in a utilitarian category where functional risk and technical performance should make peer feedback especially relevant. The results suggest a clear tension between theory and observed cross-sectional patterns: positive sentiment dominates the corpus, yet products with higher positive-share do not systematically sell more, and in this snapshot the relationship is even directionally negative. The discussion therefore emphasizes the role of measurement constraints (ceiling effects, class imbalance) and platform context (cumulative sales, exposure and credibility cues), consistent with e-WOM and reputation-signal arguments [13].

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

3.1. Sample Summary

The final dataset comprises 1,826 reviews across approximately 60 carpentry products, providing sufficient review text for sentiment estimation but a relatively small number of product-level observations for sales modeling. As displayed in Figure 2, the sentiment distribution is extremely imbalanced: roughly 94% positive, 4% neutral, and ~1% negative. This skewness is not merely a descriptive detail—it directly affects the explanatory leverage of sentiment indicators such as pos_share because high concentration near the upper bound reduces cross-product variance. In other words, when almost every product is "very positive," sentiment polarity has limited room to differentiate winners and laggards in sales. This ceiling-like pattern is mirrored by the rating behavior shown in Figure 3, where ratings cluster heavily at 4 and 5 with very few observations below 3. The average rating (4.64) and median (5.00) indicate that star ratings, like sentiment labels, suffer from a ceiling effect that constrains their usefulness as predictors in cross-sectional comparisons.

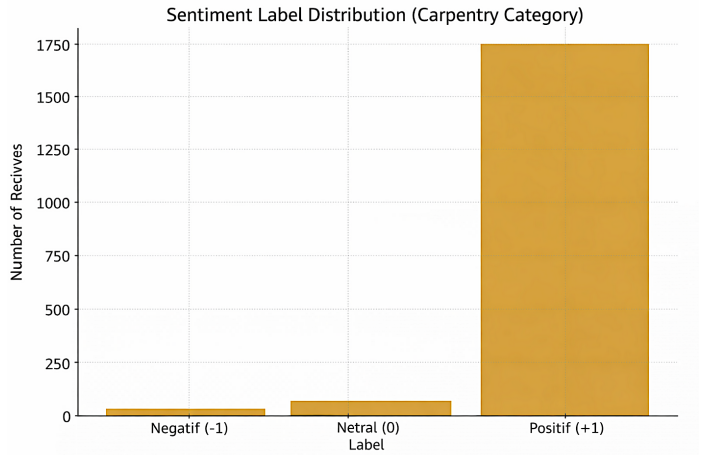

Figure 2. Sentiment Label Distribution

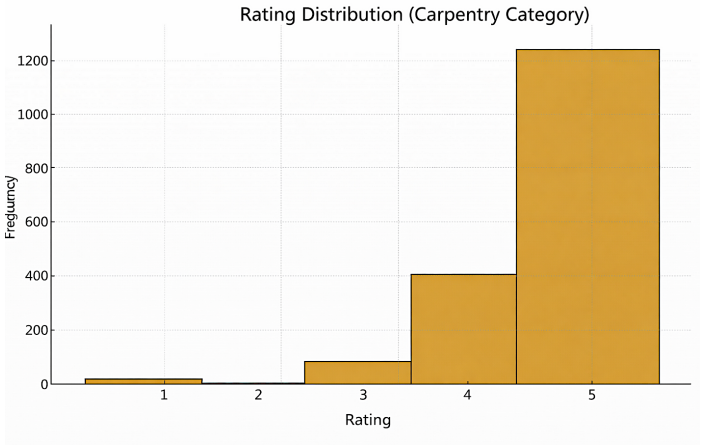

Figure 3. Rating Distribution (Histogram)

Substantively, this distribution is compatible with what buyers often do in marketplaces: they may be more likely to leave a review when satisfied, or they may rate generously unless dissatisfaction is severe. For utilitarian products like carpentry tools, another possibility is that consumers who purchase from established stores or known brands already expect acceptable performance, which can compress observed evaluative differences. The net result—visible in Figure 2 and Figure 3—is that the marketplace signal most often assumed to be decisive (valence) may be too uniform to explain sales gaps, pushing attention toward other differentiators such as exposure, store credibility, media richness (photos/video), and perceived value cues.

3.2. Correlation and Cross-Product Relationships

The bivariate plots reinforce the "high positivity, low differentiation" story while also revealing an unexpected direction of association. Figure 4 plots pos_share against log(units_sold + 1) and shows a downward-sloping trend line with a moderate negative Pearson correlation (r = -0.398). This is directionally opposite to what a straightforward reading of review-valence effects would predict from meta-analytic and experimental evidence linking positive sentiment to higher purchase intention [1], [2]. However, the negative slope becomes more interpretable once the cross-sectional structure and the bounded nature of pos_share are considered. A simple but plausible mechanism is that higher-selling products attract more heterogeneous feedback: as sales volume increases, the reviewer pool expands and includes more varied expectations, use cases, and standards, which can slightly dilute the proportion of strictly positive sentiment even while sales remain high. Conversely, low-selling products with very few reviews can easily show near-perfect positivity if only a handful of satisfied customers leave feedback. This dynamic can generate a negative correlation between pos_share and sales even if review positivity still matters at the individual decision level.

A similar pattern appears for ratings. In Figure 5, the relationship between rating_mean and log(units_sold + 1) is weaker (r = -0.166) and visually more diffuse, with no clear linear structure. The dispersion in Figure 5 is informative: products with almost identical average ratings can display widely different sales levels, suggesting that star ratings alone are insufficient to capture the competitive advantage that drives cumulative sales. This aligns with evidence that consumers attend to textual content and credibility cues,

not merely numeric ratings [3], and it also supports the Introduction's motivation for using NLP-derived sentiment rather than relying purely on rating averages.

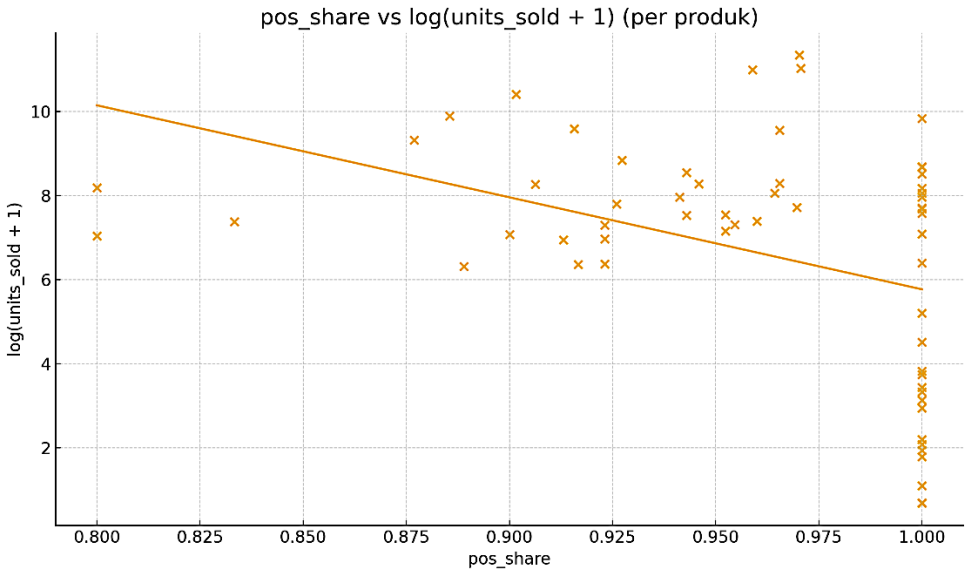

Figure 4. Scatter pos_share vs log(units_sold + 1)

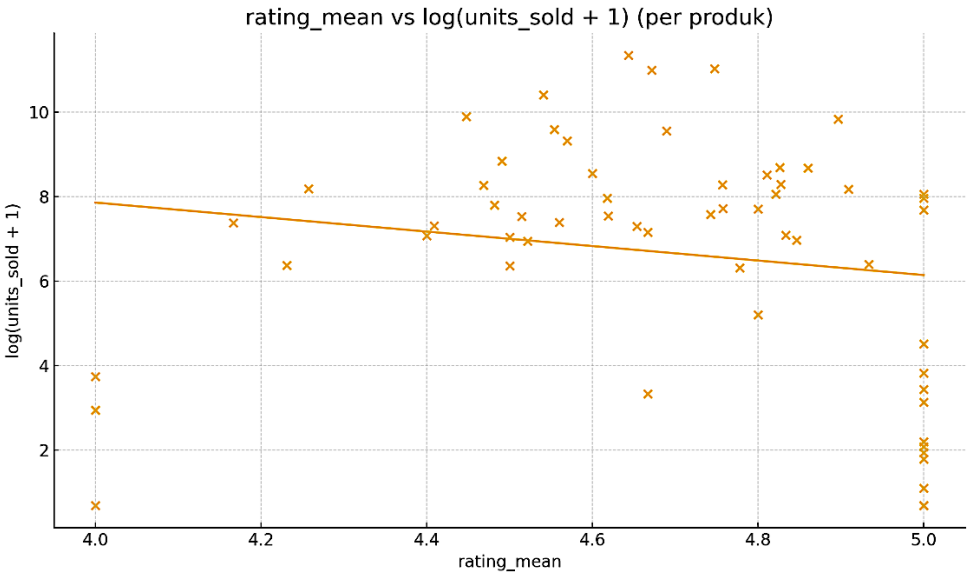

Figure 5. Scatter rating_mean vs log(units_sold + 1)

3.3. Cross-Section Regression (OLS)

To move beyond bivariate patterns, the study estimates a product-level OLS model: log(units_sold + 1) ~ pos_share + rating_mean, using the per-product aggregation (n ≈ 60). The summary results indicate that pos_share has a negative coefficient (β = -23.690, p =

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

0.004), while rating_mean is not statistically significant (β = 0.638, p = 0.667), with R^2 = 0.161. Interpreted cautiously, this model suggests that in the 2019 snapshot, variation in sentiment concentration and average rating does not dominantly explain sales variation, and the direction of the estimated association for pos_share is negative.

The key interpretive point is not that "positive sentiment reduces sales," but that pos_share is likely capturing cross-sectional structure rather than causal influence. Several mechanisms can drive the sign and significance in this setting. First, the outcome variable units_sold is cumulative and therefore entangles demand with listing age, platform exposure, and promotion history—factors not fully observed here. Second, pos_share is bounded and highly skewed, and small changes in the denominator (number of reviews) can strongly affect it, especially for products with few reviews. Third, high-selling products may accumulate more neutral/critical comments simply because more people buy them and more edge cases emerge. These points are consistent with the study's framing that the analysis tests empirical strength rather than assumes a universally positive linkage between sentiment and sales.

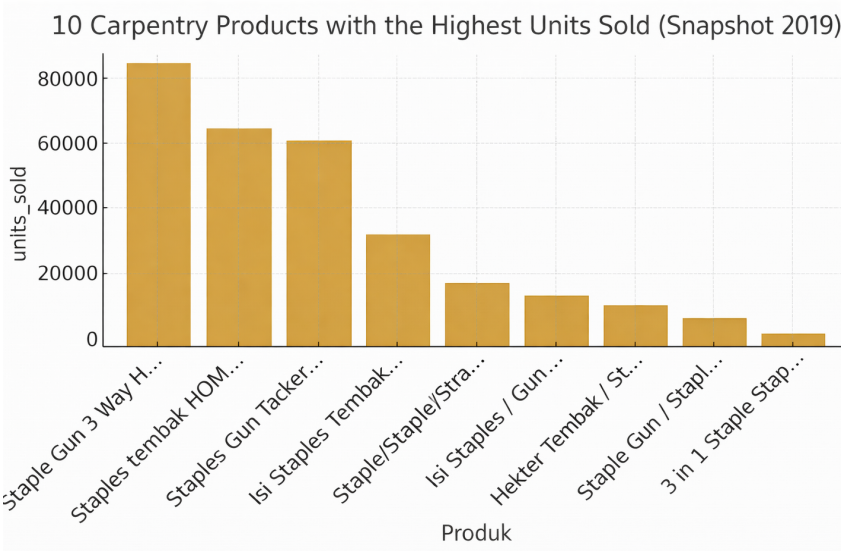

Figure 6. 10 Products with the Highest units_sold

Visual evidence from the top-selling items supports this interpretation. Figure 6, which presents the ten products with the highest units_sold, shows that several sales leaders do not exhibit uniquely high sentiment concentration relative to other products. This pattern reinforces what the regression indicates: sales dominance in this category is not

mechanically aligned with a higher share of positive polarity. In practical marketplace terms, this is where non-textual signals—such as bundling strategies, photo/video quality, seller responsiveness, shipping reliability, store reputation, and brand recognition—can plausibly outweigh minor differences in already-high sentiment, particularly when most competitors sit near the top of the rating/sentiment scale.

3.4. Discussion

The study's primary limitation is structural: the dataset has no timestamps, preventing temporal tests of whether sentiment changes precede sales changes. This matters because the broader literature often conceptualizes reviews as information that influences later decisions [1], [2], whereas a cross-sectional snapshot of cumulative units sold blends together the entire sales history. The cumulative nature of units_sold also makes results sensitive to omitted variables—price, promotion/advertising intensity, ranking exposure, listing age, and review volume—many of which are known to shape marketplace outcomes and could confound the sentiment–sales association if not controlled [12]. For robustness, it is sensible to consider outlier handling (winsorizing/trim for heavy-tailed units_sold), segmentation by product subtypes inferred from names, and alternative labeling/features. Using robust standard errors (HC) and/or quantile regression is also appropriate under heteroscedasticity and skewness, consistent with the modeling approach described earlier.

Substantively, the findings emphasize that the persuasive value of reviews in utilitarian categories depends not only on positivity but also on diagnostic content quality. Carpentry buyers often seek practical details—ease of use, durability, precision, safety, and long-term reliability—so a short "good product" comment may do little to reduce uncertainty compared to a detailed, experience-based review. This aligns with the broader claim that consumers evaluate reviews through both emotional tone and informational depth, and that attention to textual content meaningfully shapes decisions [3]. The results also fit naturally with e-WOM and trust-based interpretations: review effects are likely moderated by store credibility and platform reputation signals, where seller professionalism, complaint handling, and response speed strengthen perceived trustworthiness [13]. In that sense, a high pos_share may be necessary but not sufficient for sales conversion, particularly when almost every competing product already has high positivity.

Methodologically, Indonesian NLP work indicates that transformer-based models often outperform TF-IDF + linear baselines by capturing richer semantics, which is relevant because silver labels derived from ratings can miss nuance in the text [11], and IndoBERT can provide sentiment estimates more faithful to meaning [5]. At the same time, Tokopedia-related sentiment studies highlight class imbalance and the importance of balancing/feature choices for stability [9], [14], which is directly visible here in Figure 2. For a utilitarian category like carpentry, ABSA appears especially promising because it can map sentiment to functional aspects (power, durability, precision, package completeness) that likely drive conversion more than global positivity [7], [11]. Putting these threads together, the most consistent interpretation of this study's "weak/negative" cross-sectional pattern is that aggregate polarity alone is too coarse and too compressed (due to ceiling effects and imbalance) to explain cumulative sales differences, while marketplace outcomes are also shaped by platform exposure and credibility cues not captured in the snapshot [12], [13].

Taken as evidence, the results do not contradict the broader literature that review valence influences purchase intention [1], [2]; instead, they show that translating that micro-level effect into a product-level sales association on Tokopedia requires (i) temporal measurement, (ii) stronger commercial controls, and (iii) richer text modeling that captures diagnostic technical content rather than only polarity. This is precisely where future work can build: panel designs that align review windows with sales periods, inclusion of key commercial covariates, and ABSA/IndoBERT-based modeling to measure the specific performance attributes that matter most for carpentry tools [7], [11].

4. CONCLUSION

This study examined whether aggregated review sentiment is associated with sales performance for carpentry products on Tokopedia using a 2019 cross-sectional dataset. By modeling the relationship between the proportion of positive reviews (pos_share), average rating (rating_mean), and log(units_sold + 1), the results indicate that sentiment polarity and average ratings do not dominantly explain cross-product sales variation. The association between pos_share and sales proxy is weak and negative, while rating_mean shows no significant effect. Overall explanatory power remains limited. These findings suggest that, within a cumulative snapshot context, sales differences between products

cannot be attributed solely to sentiment polarity or star ratings. Other marketplace factors—such as pricing strategy, promotional exposure, listing age, and seller credibility—likely play a more substantial role in shaping sales outcomes. Importantly, the results do not imply that reviews are irrelevant. Rather, they indicate that aggregate polarity metrics alone are insufficient to capture the mechanisms through which reviews influence purchasing decisions. In utilitarian categories such as carpentry tools, the informational richness and technical relevance of reviews may be more influential than sentiment concentration. This study contributes by empirically testing the sentiment–sales relationship at the product level in an Indonesian marketplace context and by highlighting the limitations of cross-sectional, polarity-based modeling. Future research should incorporate temporal designs, control variables, and aspect-based sentiment approaches to better isolate the conditions under which review sentiment translates into measurable sales performance.

REFERENCES

[1] T. Chen, P. Samaranayake, X. Cen, M. Qi, and Y.-C. Lan, "The impact of online reviews on consumers' purchasing decisions: Evidence from an eye-tracking study," *Front. Psychol.*, vol. 13, Art. no. 865702, 2022.

[2] M. Ghosh, "Meta-analytic review of online purchase intention: Conceptualising the study variables," *Cogent Bus. Manag.*, vol. 11, no. 1, Art. no. 2296686, 2024.

[3] M. Kang, B. Sun, T. Liang, and H.-Y. Mao, "A study on the influence of online reviews of new products on consumers' purchase decisions: An empirical study on JD.com," *Front. Psychol.*, vol. 13, Art. no. 983060, 2022.

[4] K. Qiu and L. Zhang, "How online reviews affect purchase intention: A meta-analysis across contextual and cultural factors," *Data Inf. Manag.*, vol. 8, no. 2, Art. no. 100058, 2024.

[5] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM J. Ilm.*, vol. 14, no. 3, pp. 348–354, 2022.

[6] M. Aulia and A. Hermawan, "Analisis perbandingan algoritma SVM, Naïve Bayes, dan perceptron untuk analisis sentimen ulasan produk Tokopedia," *J. Media Inform. Budidarma*, vol. 7, no. 4, pp. 1850–1859, 2023.

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

[7] M. Z. Zainottah, R. S. Rengga, Y. S. Yustian, and I. R. Isa, "Critical sentiment analysis of Tokopedia electronic products using SVM-logistic and TF-IDF ensemble methods," *J. Artif. Intell. Eng. Appl.*, vol. 4, no. 3, pp. 2476–2482, 2025.

[8] S. Handoyo, "Purchasing in the digital age: A meta-analytical perspective on trust, risk, security, and e-WOM in e-commerce," *Heliyon*, vol. 10, no. 8, 2024.

[9] N. P. Setiawati, N. Nurmalitasari, and V. Atina, "Analisis sentimen aplikasi TikTok Tokopedia Seller Center dengan pendekatan machine learning: SVM, CNN, Naive," *Smart Comp: J. Orang Pintar Komputer*, vol. 14, no. 1, pp. 32–42, 2025.

[10] D. I. Af'idah, P. D. Anggraeni, M. Rizki, A. B. Setiawan, and S. F. Handayani, "Aspect-based sentiment analysis for Indonesian tourist attraction reviews using bidirectional long short-term memory," *JUITA: J. Inform.*, pp. 27–36, 2023.

[11] A. Andreyestha and Q. N. Azizah, "Analisa sentimen kicauan Twitter Tokopedia dengan optimalisasi data tidak seimbang menggunakan algoritma SMOTE," *Infotek: J. Inform. Teknol.*, vol. 5, no. 1, pp. 108–116, 2022.

[12] R. Meifitrah, I. Darmawan, and O. N. Pratiwi, "Sentiment analysis of Tokopedia application review to service product recommender system using neural collaborative filtering," in *IOP Conf. Ser.: Mater. Sci. Eng.*, 2020, p. 12071.

[13] O. Irnawati, G. Bayu, A. Listianto, M. Informatika, and A. Bsi Bekasi, "Metode rapid application development (RAD) pada perancangan website inventory PT. Sarana Abadi Makmur Bersama (S.A.M.B) Jakarta," 2020.

[14] H. Barus, I. N. Fajri, and Y. Pristyanto, "Sentiment classification analysis of Tokopedia reviews using TF-IDF, SMOTE, and traditional machine learning models," *J. Appl. Inform. Comput.*, vol. 9, no. 5, pp. 2552–2561, 2025.