

Aspect-Based Sentiment Analysis of Indonesian E-Commerce Product Reviews: A Unified Four-Class Sentence-Pair Approach*

Valentino Allen Prasetyo
Michael Hendrilie

Computer Science Department, Bina Nusantara University
Alam Sutera, Tangerang, Indonesia
valentino.prasetyo@binus.ac.id; michael.hendrilie@binus.ac.id

Abstract—The rapid growth of Indonesian e-commerce has produced large volumes of product reviews, creating demand for systems that extract sentiment at the aspect level. Two challenges remain underexplored: aspects expressed implicitly in informal, code-switched text, and the dominance of Aspect Term Extraction over Aspect Category Detection (ACD). This work builds an Aspect-Based Sentiment Analysis (ABSA) system for five aspects—product quality, price, shipping, seller service, and packaging—using a unified four-class formulation (negative, neutral, positive, and not-mentioned) that combines ACD and Aspect Sentiment Classification (ASC) through a sentence-pair approach following Sun et al. To isolate the contribution of contextual representation from that of the input format, all four models—TF-IDF with SVM, TF-IDF with Logistic Regression, IndoBERT, and IndoBERTweet—are trained and evaluated under an identical unified sentence-pair formulation on 1,500 manually annotated Tokopedia reviews (7,474 review–aspect pairs). The contextual models are decisively superior: their sentiment macro-F1 reaches about 0.79 against 0.35–0.54 for the lexical baselines, a gap that is statistically significant on every pairing (McNemar, $p \ll 0.05$), whereas IndoBERT and IndoBERTweet are statistically indistinguishable ($p = 0.13$)—so the hypothesis that tweet pretraining is superior is not supported. The advantage is starkest on implicit aspects, where the BERT models detect aspects with a recall of about 0.88–0.90 against 0.70–0.78 for the lexical models. Because all four models are exposed to the same aspect cue, the gap reflects contextual representation rather than differential access to the aspect information. The system is deployed as an interactive Streamlit web application that analyzes per-aspect sentiment in under one second per review.

Index Terms—*aspect-based sentiment analysis, aspect category detection, IndoBERT, IndoBERTweet, sentence-pair classification, Indonesian e-commerce*

I. INTRODUCTION

The volume of Indonesian-language product reviews has grown rapidly alongside the expansion of e-commerce platforms such as Tokopedia and Shopee, where buyers routinely document their experience across several product dimensions. Manual reading cannot keep pace with this volume, motivating automated systems that summarize opinion in a structured way. Aspect-Based Sentiment Analysis (ABSA) addresses this need by moving beyond document-level polarity to identify the specific aspects of a product and the sentiment expressed toward each. In the e-commerce

setting these aspects are concrete and recurrent—product quality, price, shipping, seller service, and packaging—which makes per-aspect analysis directly actionable for both buyers and sellers.

Two challenges remain underexplored for Indonesian ABSA. The first is implicit aspects. Indonesian e-commerce reviews are highly informal and frequently code-switched, and they often convey sentiment about an aspect without naming it: a review such as “barang sampai dalam keadaan peyok” (the item arrived dented) expresses a negative opinion about packaging even though the word packaging never appears. Lexical, bag-of-words models that rely on explicit aspect keywords tend to miss such cases.

The second challenge concerns the modeling task itself. Research on Indonesian ABSA has been dominated by Aspect Term Extraction (ATE), whereas Aspect Category Detection (ACD)—which assigns a review to a predefined set of aspect categories without requiring explicit aspect terms—remains comparatively underused, even though it is better suited to implicit aspects. Casting ACD as sentence-pair classification, in which a review is paired with an auxiliary question about each candidate aspect, lets a model relate the review to the queried aspect through cross-sentence attention. Whether this formulation actually helps, however, depends on the representation that consumes it.

This work therefore builds and compares four ABSA models—TF-IDF with SVM, TF-IDF with Logistic Regression, IndoBERT, and IndoBERTweet—on a manually annotated corpus of Indonesian Tokopedia reviews, under a unified four-class formulation that fuses ACD and Aspect Sentiment Classification (ASC). By exposing all models to the same aspect cue—the identical unified sentence-pair formulation—performance differences can be read primarily as effects of representation rather than of access to the aspect information. The study is guided by three research questions: (RQ1) how do lexical TF-IDF models compare to contextual embeddings on aspect detection and on sentiment classification; (RQ2) how effective is the sentence-pair formulation at capturing implicit aspects that contain no explicit aspect keyword; and (RQ3) how stable is each model across product categories under a cross-domain evaluation.

The contributions of this work are threefold:

* COMP6885001 — Natural Language Processing, Final Project. School of Computer Science, Bina Nusantara University, Alam Sutera Campus. Semester Ganjil (Odd) 2025/2026.

- a manually annotated four-class ABSA dataset of 1,500 Indonesian Tokopedia reviews across five aspects (7,474 review–aspect pairs);
- a unified ACD+ASC sentence-pair formulation under which all four models, spanning lexical and contextual representations, are trained and evaluated identically, exposing all models to the same aspect cue so that performance differences reflect representation rather than access to aspect information; and
- a statistical comparison (McNemar) showing that contextual models significantly outperform lexical baselines on sentiment, while tweet-pretraining yields no significant gain over general-domain pretraining.

II. RELATED WORK

ABSA decomposes opinion mining into finer-grained subtasks, principally Aspect Term Extraction (ATE), Aspect Category Detection (ACD), and Aspect Sentiment Classification (ASC) [2]. ATE locates explicit opinion-target spans in the text, whereas ACD maps a review onto a fixed inventory of aspect categories and therefore does not require the aspect to be named on the surface. This property makes ACD inherently more robust to implicit aspects, which is the regime that dominates informal Indonesian e-commerce reviews. The present study adopts the ACD view with five predefined categories and folds sentiment into the same decision through a four-class label space.

Sun et al. [6] reframed ABSA as sentence-pair classification by pairing each review with an auxiliary sentence that names the target aspect, allowing a BERT model to exploit cross-sentence attention to relate the review to the queried aspect. Yulianti and Nissa [7] adapted this construction to Indonesian product reviews and reported clear gains over single-sentence baselines, particularly for aspects that are not stated explicitly. The contextual encoders used here, IndoBERT [1] and IndoBERTtweet [5], both build on the Transformer architecture [13], whose self-attention mechanism is what enables this cross-sentence conditioning. The IndoBERT checkpoint used here is indobenchmark/indobert-base-p1, released with the IndoNLU benchmark [1]; a separate general-domain IndoBERT accompanies the IndoLEM benchmark [3]. IndoBERT is pre-trained on general-domain Indonesian text, whereas IndoBERTtweet is pre-trained on a large Indonesian Twitter corpus; the colloquial register of the latter motivates the hypothesis, tested in this work, that tweet pre-training transfers better to informal review text. As lexical baselines, TF-IDF features paired with SVM and Logistic Regression remain strong and widely used reference points for Indonesian sentiment and ABSA tasks [8], [9].

Recent comparative studies on Indonesian sentiment corroborate the broad picture adopted here. Work contrasting IndoBERT, IndoBERTtweet, and classical TF-IDF methods finds the two contextual models ahead of classical baselines but close to one another [16], and fine-tuned IndoBERT remains effective on limited-data settings [17] for both Tokopedia product reviews [18] and hotel reviews [19]. A recurring caveat is that fine-tuning BERT on small, imbalanced corpora can be unstable, with validation performance fluctuating across epochs in ways that do not reflect final quality [14]; this informs how the training dynamics reported below are interpreted. The present work

extends this line by unifying ACD and ASC in a single four-class formulation, by holding the input format constant across lexical and contextual models, and by testing the differences for statistical significance.

III. METHODOLOGY

A. Dataset and Annotation

Fig. 1 summarizes the end-to-end pipeline. The corpus comprises 1,500 Indonesian Tokopedia product reviews—1,496 of which remain after removing empty or invalid entries—drawn from three publicly available Kaggle datasets [20]–[22] and spanning three product domains: electronics, fashion, and household goods; the annotation scheme draws on the Indonesian product-review dataset PRDECT-ID [10], and the resource-documentation practice follows NusaCrowd [11]. Reviews were selected to be written in Indonesian or Indonesian–English code-switching, to contain at least five words, and to be free of exact duplicates. Two annotators labeled every review under a multi-label scheme over the five aspects; for each mentioned aspect they assigned a sentiment (positive, negative, or neutral) and a flag indicating whether the aspect was expressed implicitly, that is, without an explicit aspect keyword. Disagreements were resolved by discussion to a consensus label.

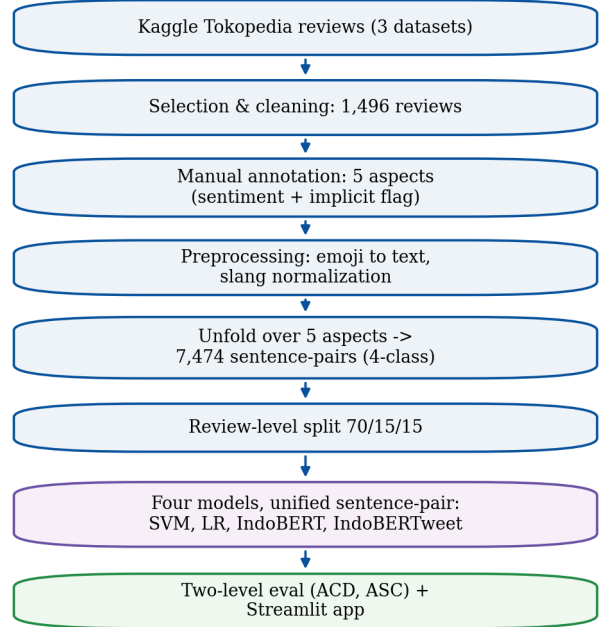


Fig. 1. End-to-end pipeline: from Kaggle reviews and manual annotation through preprocessing and aspect unfolding to the four unified sentence-pair models and the two-level evaluation.

TABLE I. ANNOTATED LABEL DISTRIBUTION ACROSS THE FIVE ASPECTS

Aspect	Mentions	Pos.	Neg.	Neu.
Product quality	1,088	484	564	40
Price	85	52	21	12
Shipping	282	175	106	1
Seller service	321	150	169	2
Packaging	179	90	86	3

^a. Counts are aspect mentions; a review may mention several aspects. Neutral is structurally rare for shipping, seller service, and packaging.

Table I summarizes the resulting distribution. The aspects are strongly imbalanced: product quality dominates, whereas price is severely under-represented (only 85 mentions, 5.7%

of reviews), and the neutral class is scarce for shipping, seller service, and packaging. These imbalances are reported openly and addressed through per-class reporting and the adaptive evaluation strategy described below, rather than concealed; no demographic attributes are used in modeling.

B. Preprocessing

Preprocessing was applied in sequence: lowercasing; conversion of emoji into descriptive text tokens, preserving their sentiment signal; removal of URLs, mentions, and hashtags; stripping of non-alphanumeric characters; reduction of characters repeated more than twice (e.g., “bagusss” → “baguss”); and slang normalization using a dictionary built on IndoCollex [12] and extended with e-commerce vocabulary (e.g., “ongkir” → “ongkos kirim”). Stemming was deliberately excluded from the pipeline: a preliminary check showed PySastrawi to be incompatible with the Python 3.14 environment used here, so the comprehensive slang-normalization step was adopted as the sole lexical-reduction mechanism instead, which proved sufficient for the aspect-classification task.

C. Problem Formulation

The task is cast as a single four-class problem that unifies ACD and ASC. Each review is unfolded into one (review, aspect) pair for all five aspects: a mentioned aspect receives its sentiment label (positive, negative, or neutral), while an absent aspect receives the label not-mentioned. This yields 7,474 pairs from the 1,496 valid reviews—1,955 mentioned and 5,519 not-mentioned. Unfolding every review across all five aspects—including those it does not mention—is what lets the model learn to detect when an aspect is absent (not-mentioned); requiring the model to evaluate every aspect of every review makes each prediction conditional on the queried aspect rather than on the overall polarity of the review, which operationalizes ACD in the sense of Sun et al. [6].

Every pair is encoded in sentence-pair form: Sentence A is the preprocessed review and Sentence B is a short auxiliary question naming the aspect (e.g., “bagaimana kualitas produk ini” for product quality). For the contextual models this is the standard [SEP]-separated BERT input [4]; for the TF-IDF models the auxiliary sentence is concatenated to the review before vectorization, an analogous construction that exposes the lexical models to the same aspect cue. The data are split 70/15/15 into training, validation, and test sets at the review level, so that all aspect pairs of a given review fall in the same partition and no information leaks across splits; the identical partition is used for every model.

D. Models

Four models are compared, all sharing the unified sentence-pair input and the identical review-level split. The two lexical models vectorize the concatenated review and auxiliary sentence with TF-IDF over unigrams and bigrams. Model 1 is a Support Vector Machine with an RBF kernel; Model 2 is L2-regularized Logistic Regression. Both use balanced class weights to counter imbalance, with no synthetic oversampling, and their hyperparameters—the regularization strength C , the TF-IDF vocabulary size, the n -gram range, and the minimum document frequency—are tuned by three-fold cross-validated grid search on macro-F1. In contrast to the conventional per-aspect baseline, a single TF-IDF classifier is trained over all aspects here so that its input format matches that of the contextual models.

Models 3 and 4 fine-tune IndoBERT (indobenchmark/indobert-base-p1) [1] and IndoBERTweet (indolem/indobertweet-base-uncased), each as a single unified four-class classifier. Both share one configuration: a maximum sequence length of 128 with only-first truncation, batch size 16, the AdamW optimizer with weight decay 0.01, a linear warmup-and-decay schedule, gradient clipping at norm 1.0, and a weighted cross-entropy loss whose class weights follow the balanced heuristic. The learning rate is selected from $\{2e-5, 3e-5, 5e-5\}$ on validation macro-F1, and training runs up to four epochs with early stopping. Holding this configuration identical across the two encoders isolates the effect of their pre-training corpus, the only intended difference between them. Consistent with [14], validation macro-F1 fluctuated across epochs in a way that did not track final test quality, so the checkpoint with the best validation macro-F1 was retained. For reproducibility, the final configuration is as follows: both lexical models use TF-IDF over unigrams and bigrams with balanced class weights and no oversampling, with the regularization strength, vocabulary size, n -gram range, and minimum document frequency fixed by the cross-validated grid search; both contextual models use the shared schedule above, and validation macro-F1 selected a learning rate of $5e-5$ for IndoBERT, with the same selection procedure applied to IndoBERTweet.

E. Adaptive Evaluation Strategy

Because the neutral sentiment is structurally rare for some aspects (Table I), sentiment is evaluated with an adaptive label space. Product quality and price retain enough neutral examples to be evaluated as ternary (positive, negative, neutral). For shipping, seller service, and packaging, where neutral is almost absent, sentiment is binary (positive, negative) and the few neutral mentions are excluded before training and evaluation. The not-mentioned class, which carries the ACD decision, is retained for all five aspects. This adaptive scheme prevents a handful of unsupported neutral cases from distorting macro-F1 and follows established practice for Indonesian e-commerce ABSA [9]. Consequently, although the formulation is four-class, three of the five aspects are evaluated over an effective three-class space (positive, negative, not-mentioned).

F. Evaluation Metrics

Because the not-mentioned class dominates the test set (819 of 1,120 pairs), overall four-class accuracy is reported only for completeness; the analysis is conducted at two levels. Aspect detection (ACD) is measured by precision, recall, and F1 on the mentioned class, capturing how well a model separates mentioned from not-mentioned pairs. Sentiment classification (ASC) is measured by macro-F1 over the effective sentiment labels, computed only on pairs whose gold label is a mention. Both ACD and ASC are averaged across the five aspects.

Three further analyses target the research questions. Detection recall is reported separately for explicit and implicit mentions to quantify the benefit of the sentence-pair formulation on aspects that carry no surface keyword (RQ2). Per-category metrics for electronics, fashion, and household goods measure stability under cross-domain evaluation, with a target spread of at most 0.10 between categories (RQ3). Pairwise McNemar tests [15] assess whether the differences between models are statistically significant rather than noise, and every model is compared against naive baselines—always

predicting not-mentioned, and predicting the majority sentiment—to confirm that performance reflects learning rather than exploitation of the class distribution. The minimum targets set for the system are an ACD mentioned-F1 of 0.75, an implicit detection recall of 0.65, and an ASC macro-F1 of 0.70.

IV. RESULTS AND DISCUSSION

A. Overall Comparison

Table II reports overall test performance. The four-class accuracies look deceptively close (0.73–0.87) because the not-mentioned class makes up 73% of the test set and is easy to predict; the picture separates once the two evaluation levels are isolated. On aspect detection the contextual models lead clearly (mentioned-F1 0.789 and 0.815 versus 0.401 and 0.446 for the lexical models), and on sentiment the gap is larger still (ASC macro-F1 about 0.79 versus 0.345 for SVM and 0.537 for LR).

The precision–recall decomposition reveals distinct error profiles. Under the unified sentence-pair format the lexical models struggle to localize which aspect is being queried: SVM is conservative, defaulting to not-mentioned and missing most genuine mentions (recall 0.387), whereas LR raises recall to 0.557 only at low precision (0.383). The contextual models instead combine moderate precision with high recall (about 0.86), recovering far more true mentions, including implicit ones analyzed below. Between the two encoders the difference is small, with IndoBERTtweet

marginally ahead on detection; whether it is meaningful is tested in the significance analysis (Fig. 2).

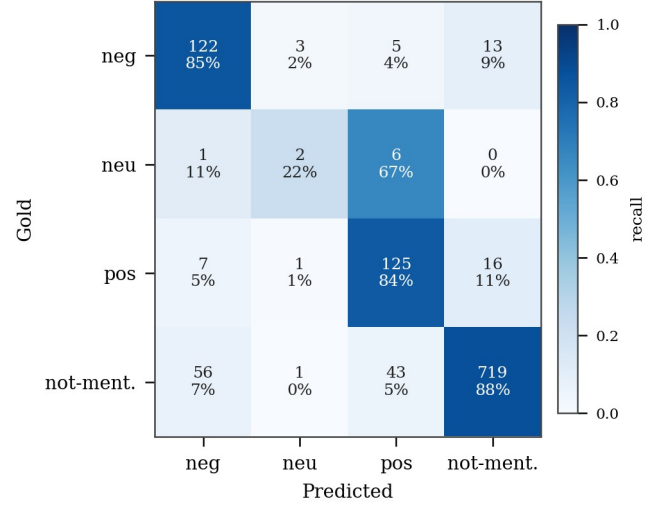


Fig. 2. Row-normalized four-class confusion matrix for IndoBERT on the test set; each cell shows the count and per-row recall. Most residual error is the rare neutral class being predicted as positive.

Both naive baselines confirm that the trained models are not exploiting the class skew. Always predicting not-mentioned reaches 0.731 accuracy but only 0.211 four-class macro-F1, while predicting the majority sentiment collapses (0.255 accuracy); all four trained models exceed both baselines by a wide margin.

TABLE II. OVERALL PERFORMANCE OF THE FOUR MODELS ON THE TEST SET (AVERAGED ACROSS ASPECTS)

Model	4-class Acc.	ACD Prec.	ACD Recall	ACD mentioned-F1	ASC macro-F1
TF-IDF + SVM	0.800	0.469	0.387	0.401	0.345
TF-IDF + LR	0.726	0.383	0.557	0.446	0.537
IndoBERT	0.864	0.730	0.862	0.789	0.794
IndoBERTtweet	0.874	0.779	0.858	0.815	0.790

^b. ACD precision/recall/F1 computed on the mentioned class; ASC macro-F1 on gold-mentioned pairs only.

B. Aspect Detection vs. Sentiment Classification

The per-aspect sentiment scores in Table III show that the contextual advantage is consistent rather than driven by a single aspect. The lexical models are competitive only on the easiest, data-rich aspect and degrade sharply elsewhere: SVM scores 0.000 on price, where only ten mentions occur in the test set, and 0.150 on packaging, whereas both BERT models stay above 0.66 on every aspect and exceed 0.91 on seller service and packaging. Sentiment is where the contextual models pull furthest ahead, because resolving polarity in informal, negation-laden Indonesian requires context that bag-of-words features cannot capture; aspect detection is comparatively easier and shows a smaller, though still clear, gap. IndoBERT and IndoBERTtweet remain close across aspects and trade places by aspect (IndoBERT leads on price, IndoBERTtweet on shipping), with no consistent winner (Fig. 3).

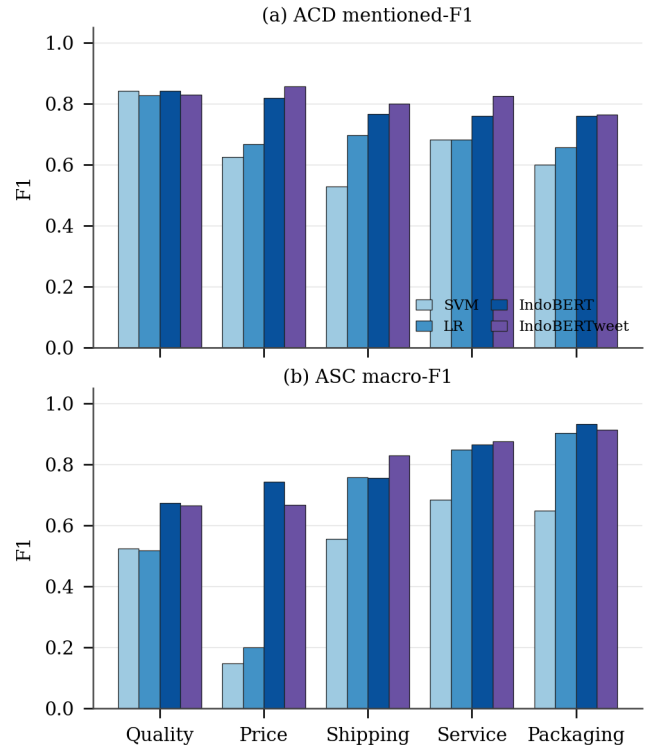


Fig. 3. Per-aspect performance of the four models: (a) ACD mentioned-F1 and (b) ASC macro-F1. Quality and price use ternary sentiment; shipping, service, and packaging use binary sentiment under the adaptive scheme.

TABLE III. PER-ASPECT SENTIMENT MACRO-F1 (ASC)

Aspect	SVM	LR	IndoBERT T	IBTwe et
Product quality	0.558	0.617	0.673	0.666
Price	0.000	0.111	0.742	0.667
Shipping	0.452	0.696	0.757	0.828
Seller service	0.566	0.712	0.866	0.876
Packaging	0.150	0.550	0.933	0.914

^c. Ternary sentiment for product quality and price; binary for shipping, seller service, and packaging (adaptive scheme).

C. Implicit Aspect Analysis

Table IV separates detection recall on explicit mentions (those containing an aspect keyword) from implicit mentions (those without), the central test of the sentence-pair formulation (RQ2). The contextual models detect implicit aspects almost as well as explicit ones—implicit recall 0.899 for IndoBERT and 0.876 for IndoBERTweet, against explicit recall 0.907 and 0.895—showing that cross-sentence attention lets them infer an aspect from context even when its keyword is absent. The lexical models are far weaker; their implicit and explicit recalls are similar (SVM 0.698/0.680, LR 0.775/0.791), but this reflects uniformly low, aspect-agnostic detection rather than genuine robustness to implicitness. All four models clear the 0.65 implicit-recall target, yet only the contextual models pair high recall with the precision needed for a strong detection F1, and their roughly 0.10–0.20 absolute recall advantage on implicit aspects is the clearest evidence that the contextual sentence-pair approach addresses the implicit-aspect gap that motivated this work (Fig. 4).

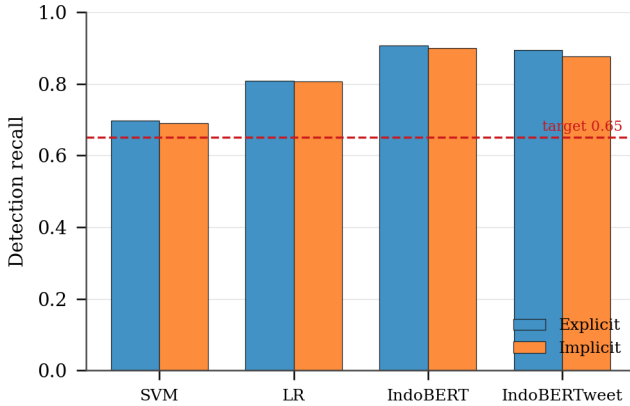


Fig. 4. Aspect-detection recall on explicit versus implicit mentions. The contextual models retain high recall on implicit mentions, whereas the lexical models stay uniformly lower; the dashed line marks the 0.65 target.

TABLE IV. DETECTION RECALL ON EXPLICIT VS. IMPLICIT MENTIONS

Model	Explicit	Implicit
TF-IDF + SVM	0.680	0.698
TF-IDF + LR	0.791	0.775
IndoBERT	0.907	0.899
IndoBERTweet	0.895	0.876

^d. Recall of detecting a gold mention; n = 172 explicit, 129 implicit.

D. Cross-Domain Evaluation

Table V reports per-domain sentiment macro-F1 with the spread (max–min) across the three categories (RQ3). Aspect detection is stable for every model: the ACD mentioned-F1 spread stays within 0.074, comfortably under the 0.10 target. Sentiment is more revealing. The contextual models are both

stable and strong (ASC spread 0.051 for IndoBERT and 0.071 for IndoBERTweet, around an absolute 0.71–0.77). Logistic Regression is unstable, its ASC spread of 0.145 exceeding the target. SVM shows a small spread (0.086) only because its sentiment quality is uniformly poor (about 0.28–0.37), so its apparent stability is not a strength. Fashion is the hardest category for every model, plausibly because its reviews blend sizing, material, and style cues that are harder to map onto a single aspect. Overall, only the contextual models generalize across domains while retaining useful sentiment quality.

TABLE V. CROSS-DOMAIN SENTIMENT MACRO-F1 AND STABILITY

Model	Elec.	Fash.	House	ΔASC	ΔACD
TF-IDF + SVM	0.362	0.282	0.368	0.086	0.029
TF-IDF + LR	0.599	0.454	0.493	0.145	0.057
IndoBERT	0.760	0.709	0.743	0.051	0.074
IndoBERTweet	0.766	0.695	0.761	0.071	0.073

^e. Per-domain values are ASC macro-F1; ΔASC and ΔACD are the max–min spreads across categories. ACD per-domain F1 is high and stable for all models.

E. Statistical Significance

Table VI gives pairwise McNemar tests on the shared 1,120-pair test set. Every comparison between a contextual model and a lexical model is significant well below the 0.001 level on both the four-class decision and the sentiment subset, confirming that the contextual advantage is not noise. The decisive comparison is IndoBERT versus IndoBERTweet: the two are statistically indistinguishable ($p = 0.13$ four-class, $p = 0.36$ on sentiment). The hypothesis that tweet-domain pre-training transfers better to informal e-commerce reviews is therefore not supported; any difference between the encoders lies within sampling noise. Because all four models are exposed to the same aspect cue, the significant lexical–contextual gaps are better explained by contextual representation than by access to the aspect information, while the null result between the two BERT models points to the pre-training corpus as a non-determining factor.

TABLE VI. PAIRWISE MCNEMAR TESTS (N = 1,120)

Comparison	p (4-cl)	p (ASC)	Sig.
SVM vs LR	2.5e-11	1.8e-4	Yes
SVM vs IndoBERT	2.4e-8	4.6e-13	Yes
SVM vs IBTweet	4.5e-11	2.1e-11	Yes
LR vs IndoBERT	6.2e-24	7.0e-8	Yes
LR vs IBTweet	6.6e-27	4.0e-6	Yes
IndoBERT vs IBTweet	0.13	0.36	No

^f. Significance assessed at $\alpha = 0.05$ on the sentiment (ASC) subset; IBTweet = IndoBERTweet.

F. Error Analysis

Table VII reports per-class F1 on the full four-class test set and locates the errors precisely. All four models handle the dominant not-mentioned class about equally well (F1 near 0.90), so the gap between lexical and contextual models comes almost entirely from the mentioned classes. Two patterns stand out. First, the neutral class is the floor for every model: with only nine neutral pairs in the test set, the lexical models never recover it at all ($F1 = 0.00$; SVM never predicts neutral), and even the contextual models reach only about 0.25. Second, the contextual advantage on negative and positive sentiment is driven by recall: IndoBERT and IndoBERTweet raise positive recall to about 0.84, against 0.56 for SVM and 0.65 for LR, while keeping precision comparable.

TABLE VII. PER-CLASS F1 ON THE FULL FOUR-CLASS TEST SET (N = 1,120; SUPPORT IN PARENTHESES)

Class	SVM	LR	IndoBERT T	IndoBERTweet
Negative (143)	0.678	0.676	0.742	0.748
Neutral (9)	0.000	0.000	0.250	0.267
Positive (149)	0.641	0.669	0.762	0.772
Not-ment. (819)	0.917	0.902	0.918	0.926

A case-level audit matches this picture. Of the 301 gold-mentioned pairs, at least one model errs on 140; the two lexical models are jointly wrong on 80 of these, and on 39 of those both contextual models are correct, while 33 errors are outright polarity reversals. The reversals cluster where surface cues mislead a bag-of-words model. In one implicit review the writer notes that the earphone arrived “peot dan terbuka” (dented and open); the quality sentiment is negative, yet both SVM and LR predict positive because no explicit quality keyword appears, whereas both contextual models read it correctly. Negation is a second recurring trap: a review praising trousers as “ga murahan” (not cheap-feeling) is gold positive, but the lexical models latch onto murahan and flip to negative, while the contextual models resolve the negation. These are exactly the implicit and negation-laden constructions that motivated this study, and they explain why the contextual lead is widest on sentiment rather than on detection.

G. Discussion

On RQ1, the contextual models dominate the lexical baselines on both subtasks under the shared sentence-pair format: ASC macro-F1 of about 0.79 versus 0.35–0.54, and ACD mentioned-F1 of 0.79–0.82 versus 0.40–0.45. The gap is widest on sentiment, where resolving polarity in informal, negation-rich Indonesian demands contextual cues that bag-of-words features lack. This is consistent with reports that fine-tuned IndoBERT stays effective for Indonesian product and hotel reviews even on limited data [18], [19], and it qualifies the strong-baseline status that TF-IDF with SVM enjoys on simpler per-aspect Shopee sentiment tasks [9]: once detection and sentiment are fused into a single aspect-conditioned decision, the lexical models can no longer keep pace.

A look at the Logistic Regression coefficients shows how the lexical models actually decide (Fig. 5). For shipping, the strongest positive feature is cepat (fast), while the strongest negative features are lama and terlambat (late or delayed), so the model keys almost entirely on explicit surface terms. It has no mechanism to handle negation or to infer an unnamed aspect, which is precisely where it breaks down on the implicit and negation-laden reviews discussed above.

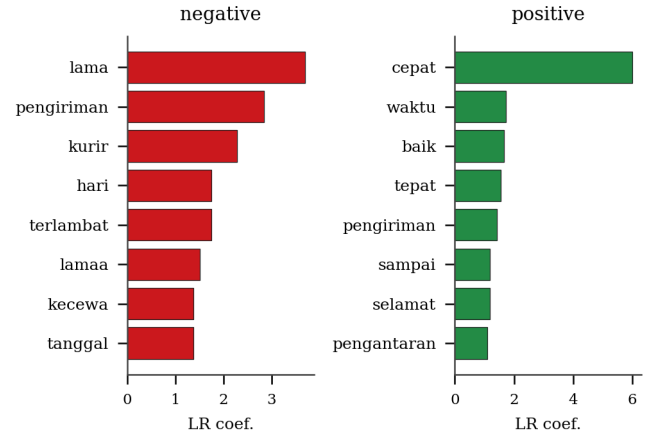


Fig. 5. Top Logistic Regression coefficients for the shipping aspect (TF-IDF features), by predicted class.

On RQ2, the sentence-pair formulation is effective precisely where it was intended to help. The contextual models lose almost nothing moving from explicit to implicit mentions (recall 0.91 to 0.90 for IndoBERT and 0.90 to 0.88 for IndoBERTweet), evidence that cross-sentence attention infers an aspect from contextual cues rather than surface keywords. The lexical models instead show flat but low recall across the explicit-implicit boundary, indicating that they never relied on aspect keywords in the first place; the auxiliary sentence adds little because a linear bag-of-words model cannot represent the interaction between the queried aspect and the review content. The implicit-aspect gap that motivated this work is thus closed by the representation, not by the format alone.

On RQ3, aspect detection generalizes across product categories for every model, but sentiment generalization is the exclusive province of the contextual models, which stay within a 0.07 spread while remaining strong; Logistic Regression drifts beyond the 0.10 target, and SVM is merely uniformly weak. Fashion is consistently the hardest category, where sizing, material, and style cues are entangled. The near-tie between IndoBERT and IndoBERTweet, statistically indistinguishable on both subtasks, mirrors recent Indonesian comparisons in which tweet-pretrained and general-domain encoders finish close together [16]; the colloquial register of e-commerce reviews is captured well enough by general-domain pre-training, and the emoji-to-text and slang-normalization steps further narrow any register advantage IndoBERTweet might have held.

Taken together, the results isolate contextual representation as the decisive factor. Because all four models are exposed to the same aspect cue and the identical review-level split, the large and uniformly significant lexical-contextual gaps are better explained by what the encoder can represent than by access to the aspect information; a fully controlled factorial ablation of format and representation is left to future work. The sentence-pair design is necessary to pose ACD as a per-aspect question, but it pays off only for a model equipped with cross-sentence attention. For practitioners building Indonesian ABSA on modest, imbalanced corpora, the implication is direct: a fine-tuned general-domain IndoBERT offers the best accuracy-stability trade-off, and the choice between tweet- and general-domain pre-training is not decisive.

V. APPLICATION

The system is delivered as an interactive web application built with Streamlit and Plotly. A user pastes a review, selects one of the trained models, and receives the predicted sentiment for each of the five aspects together with the model’s confidence, so that not-mentioned aspects and low-confidence predictions remain visible rather than hidden. The same preprocessing pipeline used in training is applied at inference, and predictions are produced in well under one second per review on commodity hardware, making the tool practical for sellers seeking aggregate, per-aspect feedback. Fig. 6 shows the interface.

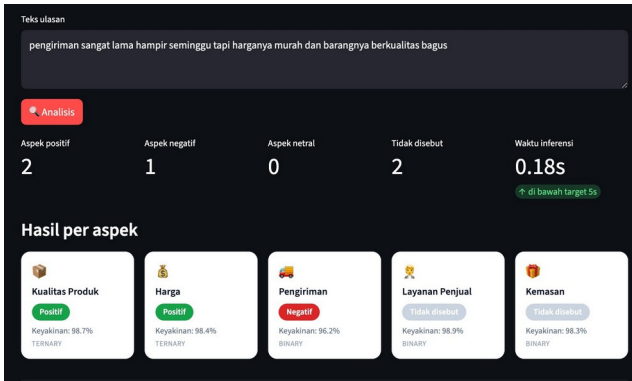


Fig. 6. The Streamlit web application: predicted sentiment and confidence for each of the five aspects of an input review.

VI. LIMITATIONS

Several limitations qualify these results. First, inter-annotator agreement is not quantified; annotation quality rests on dual labeling with consensus resolution rather than a reported coefficient. Second, the corpus is drawn solely from Tokopedia, so generalization to other platforms is untested. Third, although all models share the unified sentence-pair input, a linear bag-of-words model cannot exploit the aspect cue the way cross-sentence attention does, so representation and input format are not fully disentangled; a controlled ablation is left to future work. Fourth, under the adaptive scheme the neutral class is dropped for shipping, seller service, and packaging, so the four-class space is effectively three-class for those aspects. Finally, the price aspect is severely under-represented—ten mentions in the test set—so its per-aspect figures are indicative only.

VII. CONCLUSION AND FUTURE WORK

This work framed Indonesian e-commerce ABSA as a unified four-class problem that fuses aspect category detection and sentiment classification, and compared lexical and contextual models under an identical sentence-pair formulation. The contextual models met every target—ACD mentioned-F1 of 0.79–0.82 against the 0.75 goal, ASC macro-F1 near 0.79 against 0.70, and implicit-aspect detection recall of 0.88–0.90 against 0.65—whereas the lexical models fell short on both detection and sentiment. In answer to the research questions: contextual embeddings clearly outperform TF-IDF (RQ1); the sentence-pair formulation captures implicit aspects effectively, but only when paired with a contextual encoder (RQ2); and only the contextual models generalize across product categories while retaining sentiment quality (RQ3). Tweet-domain pre-training conferred no significant advantage over general-domain pre-training.

Several directions follow. The corpus is modest and imbalanced—price in particular is scarce—so expanding and balancing it across more product categories would sharpen the cross-domain picture. Few-shot and prompt-based methods are a promising route for the under-represented aspects and the rare neutral class. Finally, a controlled ablation that independently varies the input format and the representation would quantify how much each contributes, complementing the format-controlled comparison reported here.

REFERENCES

- [1] B. Wilie et al., “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” in Proc. AACL-IJCNLP, 2020.
- [2] M. Pontiki et al., “SemEval-2015 Task 12: Aspect Based Sentiment Analysis,” in Proc. SemEval, 2015.
- [3] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” in Proc. COLING, 2020.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805, 2018.
- [5] F. Koto, J. H. Lau, and T. Baldwin, “IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Media Bias Analysis,” in Proc. EMNLP, 2021.
- [6] C. Sun, L. Huang, and X. Qiu, “Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence,” in Proc. NAACL-HLT, 2019.
- [7] E. Yulianti and F. K. Nissa, “Aspect-Based Sentiment Analysis of Indonesian Product Reviews Using Sentence-Pair Classification,” Int. J. Adv. Comput. Sci. Appl., 2022.
- [8] A. H. Jazuli, S. Adhy, and P. W. Wirawan, “Aspect-Based Sentiment Analysis on Indonesian E-Commerce Reviews,” in Proc. ICACIS, 2022.
- [9] N. Wulandari, M. A. Fauzi, and B. D. Setiawan, “Sentiment Analysis of Shopee Reviews Using SVM and TF-IDF,” JTIK, 2021.
- [10] H. T. T. Do et al., “PRDECT-ID: Indonesian Product Reviews Dataset for Aspect-Based Sentiment Analysis,” Data in Brief, 2022.
- [11] S. Cahyawijaya et al., “NusaCrowd: Open Source Initiative for Indonesian NLP Resources,” in Proc. ACL, 2023.
- [12] H. A. Wibowo et al., “IndoCollex: A Testbed for Morphological Feature Extraction of Indonesian Colloquial Words,” in ACL Anthology, 2021.
- [13] A. Vaswani et al., “Attention Is All You Need,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.
- [14] M. Mosbach, M. Andriushchenko, and D. Klakow, “On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines,” in Proc. ICLR, 2021.
- [15] T. G. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms,” Neural Comput., vol. 10, no. 7, pp. 1895–1923, 1998.
- [16] F. Alvin and N. A. S. Winarsih, “Perbandingan Kinerja Model IndoBERT, IndoBERTweet, dan Algoritma Klasik pada Analisis Sentimen Isu Indonesia Gelap,” BITS, vol. 7, no. 3, pp. 1601–1613, 2025.
- [17] M. A. Mayzaroh, D. F. Ningsih, N. Destriani, and M. C. T. Manullang, “Benchmarking PyCaret AutoML Against IndoBERT Fine-Tuning for Sentiment Analysis on Indonesian IKN Twitter Data,” arXiv:2604.25392, 2026.
- [18] K. Wau, “Application of Fine-Tuned IndoBERT for Sentiment Classification of Local Product Reviews on Tokopedia Marketplace with Limited Dataset,” JAIEA, 2025.
- [19] S. Apriliani, A. Erfina, and C. Warman, “Fine-Tuned IndoBERT for Aspect-Based Sentiment Analysis of Indonesian Five-Star Hotel Reviews,” J. Sisfokom, vol. 14, no. 4, pp. 437–445, 2025.
- [20] farhan999, “Tokopedia Product Reviews,” Kaggle dataset. [Online]. Available: <https://www.kaggle.com/datasets/farhan999/tokopedia-product-reviews>
- [21] silviarmargareta, “Dataset Review,” Kaggle dataset. [Online]. Available: <https://www.kaggle.com/datasets/silviarmargareta/dataset-review>

- [22] salmanabdu, “Tokopedia Product Reviews 2025,” Kaggle dataset, 2025. [Online]. Available: <https://www.kaggle.com/datasets/salmanabdu/tokopedia-product-reviews-2025>

APPENDIX

A. Team Contribution Statement

Valentino Allen Prasetyo led the project scoping and research-question design (RQ1–RQ3), the text preprocessing pipeline, the unified four-class sentence-pair problem formulation, the fine-tuning of IndoBERT and IndoBERTweet (Models 3 and 4), the complete evaluation suite (aspect-detection and sentiment metrics, implicit-aspect analysis, cross-domain evaluation, and the McNemar significance tests), the error analysis, the demonstration video, and repository and version-control management. Michael Hendrilie led the dataset sourcing and curation from the three public Kaggle Tokopedia collections, the data cleaning and review selection, the TF-IDF Support Vector Machine and Logistic Regression baselines (Models 1 and 2), the Streamlit web application, and the presentation slides. Both authors jointly performed the manual annotation of all 1,500 reviews—including independent dual labeling and consensus resolution of disagreements—and jointly wrote this report and IEEE paper. The team coordinated through regular discussions and tracked all work using Git and GitHub version control.

B. Code and Data Availability

The full source code, the manually annotated dataset, and the trained-model artifacts are publicly available at: **<https://github.com/ValentinoAllen/ABSA-application>**

A demonstration video of the application is available at: **<https://youtu.be/9NPseZAbvTc>**