

# Fake Review Detection on Indonesian E-Commerce Platforms Using a Hybrid CNN-BiLSTM-Attention Architecture Optimized by Differential Evolution

Indu Indah Purnomo<sup>1</sup>, Rina Alfah<sup>2</sup>, Hayati Noor<sup>3</sup>

<sup>1,2,3</sup> Faculty of Information Technology, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari, Banjarmasin, Indonesia  
[induindahpurnomo@gmail.com](mailto:induindahpurnomo@gmail.com)<sup>1</sup>, [rina.alfah.05@gmail.com](mailto:rina.alfah.05@gmail.com)<sup>2</sup>, [hayatinoor@uniska-bjm.ac.id](mailto:hayatinoor@uniska-bjm.ac.id)<sup>3</sup>

## Article Info

### Article history:

Received 2026-08-02

Revised 2026-10-17

Accepted 2026-10-19

### Keyword:

*BiLSTM,*  
*CNN,*  
*Cross-dataset.*

## ABSTRACT

Fake reviews on e-commerce platforms threaten consumer trust and harm local micro, small, and medium enterprises (MSMEs). This study develops a fake review detection system using a hybrid deep learning architecture—Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and an Attention mechanism—whose hyperparameters are optimized by Differential Evolution (DE). A two-phase design is proposed: Phase 1 validates the architecture on an English benchmark (Kaggle Fake Reviews, 40,432 reviews), and Phase 2 generalizes it to an Indonesian e-commerce corpus (38,825 cleaned reviews from Tokopedia and Shopee) labeled automatically via Latent Dirichlet Allocation (LDA). Because no Indonesian corpus records review authorship, the validity of the induced labels is treated as a research question rather than an assumption: the topic-to-class mapping is anchored to blind annotation by two independent annotators, and agreement is quantified with Cohen's Kappa on a held-out half of the annotated sample. The proposed model was compared against four baselines (Standalone LSTM, CNN+PSO, BiLSTM+GA, and Standalone DE) across six metrics over independent runs, with One-Way ANOVA and Tukey HSD post-hoc tests. The proposed model achieved competitive accuracy with no statistically significant difference from the strongest deep-learning baselines. Accuracy on the Indonesian phase should be read with caution, since labels induced from word distributions are predictable from the same textual regularities the classifier learns, while Standalone DE without a hybrid architecture performed significantly worse. The main contributions are an openly documented labeling and validation procedure for Indonesian OR/CG data, an equal-budget comparison of DE against PSO and GA, and a cross-dataset generalization analysis situated in the MSME context.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

## I. INTRODUCTION

This study requires an operational definition, because the term “fake review” is used loosely in the literature. Following Mohawesh et al. [1] and Kumar et al. [2], we define a fake review as a review whose content was not produced by a genuine purchaser reporting an authentic consumption experience. Two classes follow from this definition. An Original (OR) review is written by a real customer about a product actually obtained. A Computer-Generated (CG) review is produced by an automated system, or written to a template by a party who did not consume the product, in order to influence consumer perception. The

definition is deliberately narrower than the everyday usage of “fake”: it excludes reviews that are genuine but inaccurate, exaggerated, or written in bad faith by an actual purchaser, since those cannot be separated from authentic reviews on textual evidence alone.

This definition also fixes what the present study can and cannot claim. Authorship is not observable in any of the corpora used here; no dataset of Indonesian e-commerce reviews records whether a review was machine-generated. Any label we assign is therefore an inference from textual and distributional evidence, not a ground truth. Section III-I reports the consequences of this limitation openly rather than treating the assigned labels as if they were verified.

Fake review detection has been extensively studied using machine learning and deep learning approaches [3]. Methods based on Long Short-Term Memory (LSTM) have proven capable of capturing the linguistic patterns that distinguish genuine from fabricated reviews [4], [15], while combining CNN with BiLSTM improves accuracy by extracting local features alongside sequential dependencies [5]. Adding an attention mechanism further strengthens the model's ability to weight the most informative parts of the text [6].

Singh et al. [7] demonstrated that a hybrid CNN-LSTM-Attention architecture optimized by Differential Evolution (DE) achieves 96% accuracy in detecting fabricated content, outperforming Particle Swarm Optimization (PSO) [8], Genetic Algorithm (GA) [9], and standalone deep learning models. DE is a population-based optimization algorithm that is effective for tuning neural network hyperparameters [10]. That study, however, focused on English-language fake news and did not address Indonesian fake reviews, which exhibit distinct linguistic characteristics: heavy use of slang, code-mixing, informal abbreviations, and culturally specific emotional expressions. These informal and non-standard characteristics have consistently posed challenges in Indonesian e-commerce text analysis [16], [17], both for classical approaches and for transformer-based models such as IndoBERT [18], [19].

The principal obstacle to Indonesian fake review detection is the absence of a publicly available dataset labeled as Original or Computer-Generated (OR/CG) [11]. This study addresses that gap by integrating automatic labeling based on Latent Dirichlet Allocation (LDA) validated by human annotators, DE-based optimization, and a standardized cross-dataset validation procedure.

#### A. Related Work

Research on fake review detection has evolved through three successive waves. The first wave relied on classical machine learning with manual feature engineering. Barbado et al. [24] designed a detection framework based on user behavioral features and review metadata for online consumer electronics retailers, showing that non-textual features such as review frequency and account history carry strong discriminative power. A comparable approach was applied by Purifyregalia et al. [11] to Shopee Indonesia reviews using Support Vector Machine and Random Forest with labels derived from LDA. The principal weakness of this wave is its dependence on hand-crafted features that transfer poorly across domains and languages.

The second wave employed deep learning to learn text representations automatically. In their comprehensive survey, Mohawesh et al. [1] documented a decisive shift from manual features toward distributed representations. Padalko et al. [4] showed that LSTM captures long-range dependencies in fake news text, whereas Bathla et al. [5] combined aspect extraction with deep learning for aspect-based fake review detection. Hybrid architectures subsequently proved superior: Srikanth et al. [15] combined CNN and BiLSTM with a modified feature attention mechanism, and Jian et al. [6] proposed SA-Bi-LSTM, which integrates self-attention with BiLSTM. The rationale behind

these combinations is consistent: CNN excels at capturing local n-gram patterns such as repeated superlative phrases, BiLSTM encodes bidirectional sequential context, and attention assigns weights to the most discriminative tokens.

The third wave introduced metaheuristic optimization for hyperparameter tuning, motivated by the strong sensitivity of deep learning architectures to configuration choices. Deshai and Rao [8] applied adaptive PSO to CNN, while Al Bataineh et al. [9] optimized Bi-LSTM using a genetic algorithm. Baiocchi et al. [10] systematically demonstrated that DE serves as an efficient optimization engine for neural networks, and Sen et al. [23] employed it to tune transformer hyperparameters. Jiao et al. [22] concluded that evolutionary algorithms perform well on non-differentiable and multimodal search spaces—precisely the characteristics of neural network hyperparameter spaces. Singh et al. [7] unified all three waves through a DE-optimized CNN-LSTM-Attention architecture, which serves as the principal reference for the present study.

In the Indonesian context, natural language processing research has concentrated on sentiment analysis rather than fake review detection. Romadhony et al. [16] released a large-scale Indonesian product review dataset for sentiment classification, Sutoyo et al. [13] compiled PRDECT-ID for emotion classification, and Lin and Nuha [17] proposed a hybrid strategy across several Indonesian datasets. On the representation side, Koto et al. [18] released IndoBERT as the reference pre-trained model for Indonesian, followed by Murfi et al. [19], who combined BERT with CNN and RNN, and Ahmadian et al. [25], who developed hybrid models for emotion classification. To the authors' knowledge, however, no open and validated OR/CG-labeled Indonesian e-commerce review dataset exists—the gap this study fills.

Narrowing to fake review detection on Indonesian e-commerce specifically, the literature is thin and methodologically uniform. Purifyregalia et al. [11] classified Shopee reviews with Support Vector Machine and Random Forest, deriving labels from Latent Dirichlet Allocation rather than from verified authorship. To the authors' knowledge, no published study on Indonesian fake review detection has combined deep learning with an attention mechanism, and none has applied evolutionary hyperparameter optimization to this task. Equally important, no published study has validated the labels it used against human annotators, which means the reported accuracies in this subfield rest on label-generation procedures that have not themselves been tested.

The gap this study addresses is therefore twofold. The first is architectural: a hybrid CNN-BiLSTM-Attention model optimized by Differential Evolution has been evaluated on English fake news [7] but not on Indonesian fake reviews. The second, and in retrospect the more consequential, is evidential: whether automatic topic-based labeling can produce labels that a human annotator would agree with. The present work treats the second question as a research question in its own right rather than as a preliminary step.

## B. Research Contributions

This study offers four contributions. First, an openly specified procedure for inducing and validating OR/CG labels on Indonesian e-commerce reviews, including the topic-to-class anchoring step that comparable studies leave implicit. We describe the resulting resource as a labeled corpus accompanied by an agreement measurement, not as a gold-standard dataset, since authorship is never observed. Second, the implementation and evaluation of a DE-optimized hybrid CNN-BiLSTM-Attention architecture in the fake review domain, which had previously been tested only on English fake news. Third, a cross-dataset comparison run through an identical pipeline, together with an explicit account of why the observed differences cannot be attributed to language. Fourth, transparent reporting of the limitations of automatic LDA labeling, including evidence of label circularity that is rarely disclosed in comparable literature.

The research questions are: (1) how to design and implement a DE-optimized hybrid CNN-BiLSTM-Attention architecture; (2) how effective automatic LDA labeling is in producing a valid Indonesian fake review dataset; (3) how model performance compares between the English dataset (Phase 1) and the Indonesian dataset (Phase 2); and (4) how the proposed model compares against four baselines.

## II. RESEARCH METHOD

This study adopts a two-phase computational experimental approach. Phase 1 validates the architecture on an international benchmark dataset; Phase 2 generalizes the model to the Indonesian context through an LDA labeling pipeline. Comparing the results of both phases yields the cross-dataset analysis that constitutes the novelty of this work. The overall research flow is presented in Fig. 1.

### A. Datasets and Preprocessing

Phase 1 uses the Kaggle Fake Reviews Dataset, comprising 40,432 OR/CG-labeled reviews [12]. Phase 2 combines three sources of Indonesian product reviews (Tokopedia 2019, Shopee/Tokopedia, and PRDECT-ID [13]), totaling 96,930 raw reviews. Indonesian preprocessing includes filtering corrupted text and machine-translated content, language filtering based on marker-word ratios, slang normalization, and stemming with Sastrawi [20], yielding 38,825 clean reviews.

The Indonesian corpus is assembled from three public sources rather than collected by the authors, which constrains what can be reported about its provenance. The first source is a 2019 Tokopedia product review dump; the second is a combined Shopee and Tokopedia file; the third is PRDECT-ID [13]. Marketplace provenance can be established for the first and third sources but not the second, because that file merges reviews from two marketplaces without a column identifying the origin of each row. Rows from that file are therefore excluded from any analysis that depends on marketplace identity.

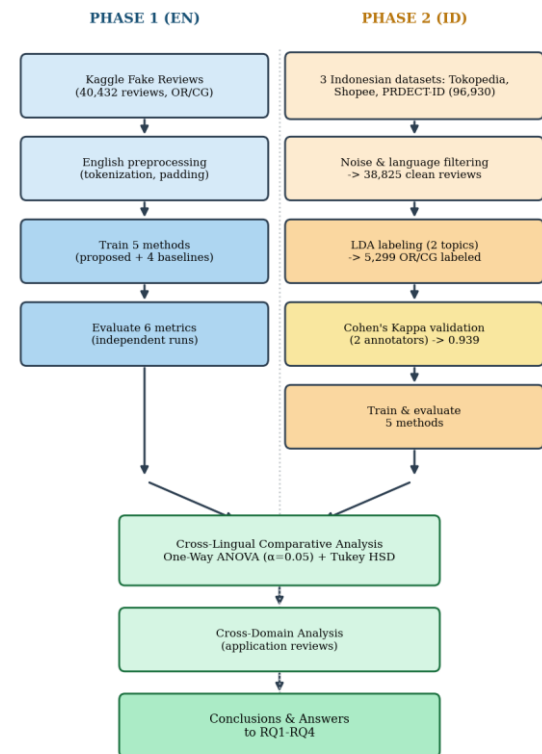


Fig. 1. Two-phase research flow: architecture validation on the English dataset (Phase 1) and generalization to Indonesian with LDA labeling (Phase 2).

Two attributes that reviewers of this manuscript reasonably expect are unavailable. Neither the collection period nor the product category is recorded consistently across the three sources, and none of them carries product or seller identifiers. The absence of seller and product identifiers has a direct methodological consequence: we cannot guarantee that reviews of the same product or from the same seller do not appear on both sides of a train-test split. Near-duplicate filtering, described in Section II-B, mitigates the most obvious form of this leakage but does not eliminate it.

### B. LDA Labeling and Validation

Because no Indonesian corpus records review authorship, labels were induced with Latent Dirichlet Allocation [21] over a bag-of-words representation with two topics. Documents whose dominant topic probability exceeded 0.6 were retained and assigned to the class of that topic; ambiguous documents were discarded. The mapping from topic index to the OR and CG classes is the step that carries the entire evidential weight of the procedure, and it is not determined by LDA itself.

This point deserves emphasis because it is easily glossed over. Latent Dirichlet Allocation is a topic model: it partitions documents by word co-occurrence and has no notion of authorship or authenticity. Nothing in the algorithm identifies which of its two topics, if either, corresponds to machine-generated text. The mapping must be supplied from

outside the model, and the justification for that mapping is what determines whether the resulting labels mean anything.

We originally mapped topics using a lexical heuristic, assigning the topic richer in superlative and emotional vocabulary to the CG class, on the reasoning of Kumar et al. [2] that fabricated reviews carry higher sentiment extremity and less product-specific detail. Re-examination showed this heuristic to be unsafe. The two topics differed by a single word under the original word list, and the list itself contained concrete attributes such as “fast” and “cheap” that describe delivery and price rather than exaggerated praise. The mapping was therefore decided by vocabulary that the underlying argument does not support.

The revised procedure anchors the mapping to human annotation instead. Annotated reviews are split into disjoint halves: one half determines which topic corresponds to the CG class, the other validates the resulting labels. The split matters, because using the same annotations for both purposes would make the validation circular. The mapping is accepted only when the agreement on the anchoring half is significantly better than chance under a binomial test; otherwise the pipeline halts and reports that the labels are not usable.

Label validity is assessed by two annotators working independently on a random sample drawn from the frozen dataset. The annotation file carries no LDA output, so annotators are blind to the labels under test. Agreement is quantified with Cohen's Kappa [14]. We stress what this statistic can and cannot establish: Kappa measures whether two annotators applied the same criteria consistently, not whether those criteria track authorship. High agreement between annotators, and even high agreement between annotators and the induced labels, is compatible with both parties responding to writing style rather than to authenticity. For this reason we do not describe the resulting resource as a gold-standard dataset, and we report the labeling procedure as a research question whose answer is reported in Section III-I rather than as a solved preliminary.

### C. Proposed Model Architecture

The CNN-BiLSTM-Attention+DE model consists of five sequential layers, as illustrated in Fig. 2: (1) an embedding layer; (2) a multi-kernel CNN with kernel sizes of 3, 4, and 5 for local feature extraction; (3) a BiLSTM layer for bidirectional dependencies [5]; (4) an additive attention layer that weights informative features [6]; and (5) a fully connected layer. Hyperparameters are optimized by DE through the mutation operation  $V_i = X_{r1} + F \cdot (X_{r2} - X_{r3})$ , crossover, and selection based on the fitness function  $F(X) = w_1 \cdot \text{Acc} - w_2 \cdot \text{Err} + w_3 \cdot \text{Rob}$  ( $w_1=0.5$ ;  $w_2=0.3$ ;  $w_3=0.2$ ), with  $F=0.7$  and  $CR=0.5$  [7].

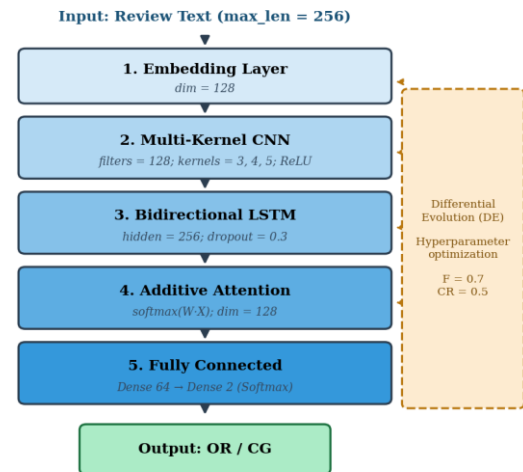


Fig. 2. Architecture of the proposed CNN-BiLSTM-Attention model with Differential Evolution (DE) hyperparameter optimization.

The hyperparameter search space optimized by DE is summarized in Table I. Each individual in the DE population represents one architectural configuration encoded as a real-valued vector, which is rounded to the nearest discrete value during evaluation. A population of 30 is evolved over 30 generations, producing 900 fitness evaluations per experiment. An identical evaluation budget was allocated to PSO and GA to ensure a fair comparison—an aspect frequently overlooked in comparative metaheuristic studies.

TABLE I  
HYPERPARAMETER SEARCH SPACE FOR DE OPTIMIZATION

| Hyperparameter        | Value Range | Type           |
|-----------------------|-------------|----------------|
| Embedding dimension   | 64 – 256    | Discrete       |
| Number of CNN filters | 64 – 256    | Discrete       |
| BiLSTM units          | 64 – 256    | Discrete       |
| Dropout rate          | 0.2 – 0.5   | Continuous     |
| Learning rate         | 1e-4 – 1e-2 | Log-continuous |
| Batch size            | 32 – 128    | Discrete       |

The maximized fitness function combines three components: validation accuracy (Acc), error rate (Err), and robustness (Rob), the latter measured as the inverse of the standard deviation of accuracy across folds. The formulation  $F(X) = w_1 \cdot \text{Acc} - w_2 \cdot \text{Err} + w_3 \cdot \text{Rob}$  with weights of 0.5, 0.3, and 0.2 follows Singh et al. [7] so that results remain directly comparable with the reference study. The robustness component was deliberately included to prevent DE from selecting configurations that happen to excel on a single data split yet remain unstable overall.

The optimization procedure is specified as follows so that it can be reproduced. Each individual encodes the six hyperparameters of Table I as a real-valued vector in the unit hypercube, decoded to the stated ranges at evaluation time; the learning rate is searched on a logarithmic scale because its range spans two orders of magnitude. Mutation follows the classical DE/rand/1 scheme,  $V_i = X_{r1} + F(X_{r2} - X_{r3})$ , with three distinct randomly chosen individuals and scale factor  $F = 0.7$ . Binomial crossover with rate  $CR = 0.5$  combines the donor and target vectors, with one dimension

forced to come from the donor so that the trial vector always differs from its parent. Selection is greedy: the trial vector replaces the target only if its fitness is higher.

The stopping criterion is a fixed generation count rather than a convergence threshold, because a fixed budget is what makes the comparison against PSO and GA interpretable. Fitness is computed on the validation split only. The test split is untouched during optimization and is used exactly once per run, after the final model has been trained with the selected configuration; no model selection decision anywhere in the pipeline consults it.

#### D. Implementation Environment

All experiments were implemented in Python 3.11 using TensorFlow/Keras for modeling, Sastrawi for Indonesian stemming [20], Gensim and scikit-learn for LDA labeling [21], and statsmodels for ANOVA and Tukey HSD testing. Computation was performed in a Kaggle Notebook environment with GPU acceleration. To ensure reproducibility, a random seed was fixed for every run and all experimental configurations were stored in separate files. Data were split at a 70:15:15 ratio for training, validation, and testing, with class stratification to preserve the OR and CG proportions in each subset.

Class imbalance in the Indonesian dataset (3,746 CG versus 1,553 OR) was addressed through automatic class weighting computed inversely to class frequency. Without this mechanism, models tend to collapse toward the majority class—a phenomenon observed precisely in the Standalone DE baseline and discussed in Section III.

#### E. Baselines, Metrics, and Statistical Testing

The proposed model was compared against four baselines: Standalone LSTM [4], CNN+PSO [8], BiLSTM+GA [9], and Standalone DE [10]. To ensure a fair comparison, all evolutionary optimization algorithms (DE, PSO, GA) received an equivalent fitness evaluation budget through identical population sizes and iteration counts. Evaluation employed six metrics: Accuracy, Precision, Recall, F1-Score, Mean Error Rate (MER), and Convergence Rate. Differences were tested using One-Way ANOVA at  $\alpha=0.05$ , followed by Tukey HSD post-hoc testing.

The fourth baseline requires clarification, since Differential Evolution is an optimizer and not a classifier, and so is not directly comparable to CNN or LSTM. In the Standalone DE condition, DE optimizes the hyperparameters of a single fully connected network over an averaged embedding of the review text, with no convolutional, recurrent, or attention component. The condition is therefore an ablation of the architecture rather than a competing method: it isolates what evolutionary optimization contributes when it is not paired with a sequence model. Reporting it alongside the other baselines is what allows the contribution of the hybrid architecture to be separated from the contribution of the optimizer.

Comparability across the three evolutionary optimizers is enforced by construction. DE, PSO, and GA receive identical population sizes and iteration counts, hence an identical number of fitness evaluations per run, and each fitness

evaluation trains for the same number of epochs on the same subset. The experiment script verifies this equality at start-up and refuses to proceed silently if the three budgets differ. We note that the budget used here is smaller than the one specified in the original research plan. The reduction was forced by the computational cost of the full protocol, and we report the actual budget rather than the planned one; what the comparison requires is that the three optimizers be treated equally, which holds regardless of the absolute budget.

These metrics capture complementary aspects of performance: Accuracy measures the proportion of correct predictions, Precision the exactness of fake-class predictions, Recall the ability to retrieve all fake reviews, F1-Score the harmonic mean of Precision and Recall, MER the complement of accuracy, and Convergence Rate the pace of fitness improvement across generations for metaheuristic-based methods. In fake review detection, Precision carries substantial practical weight because incorrectly flagging genuine reviews as fake (false positives) harms honest sellers, while Recall matters so that manipulative reviews do not escape detection. F1-Score is therefore treated as the primary metric throughout the discussion.

Each method was executed repeatedly with independent initializations to obtain a distribution of performance rather than a single point estimate. Differences among methods were tested with One-Way ANOVA at a significance level of 0.05; where the result was significant, the Tukey Honestly Significant Difference test identified which specific pairs of methods genuinely differ. This two-stage procedure is essential because ANOVA establishes only that a difference exists, without indicating where it lies.

### III. RESULTS AND DISCUSSION

#### A. Method Comparison in Phase 1 (English)

Table II presents the performance of all five methods on the English dataset. BiLSTM+GA attained the highest accuracy (0.9224), followed by the proposed model (0.9128). One-Way ANOVA indicated a significant difference among methods ( $F=8.33$ ;  $p=0.003$ ).

TABLE II  
PHASE 1 COMPARISON RESULTS (ENGLISH)

| Method                     | Acc    | F1     | MER    | CR     |
|----------------------------|--------|--------|--------|--------|
| Proposed CNN-BiLSTM-Att+DE | 0.9128 | 0.9136 | 0.0872 | 0.0036 |
| Standalone LSTM            | 0.9089 | 0.9093 | 0.0911 | —      |
| CNN+PSO                    | 0.9000 | 0.9017 | 0.1000 | 0.0019 |
| BiLSTM+GA                  | 0.9224 | 0.9228 | 0.0776 | 0.0017 |
| Standalone DE              | 0.8478 | 0.8580 | 0.1522 | 0.0975 |

Values are averaged across runs. The Tukey HSD test showed that the proposed model differs significantly only from Standalone DE ( $p=0.008$ ); against BiLSTM+GA, LSTM, and CNN+PSO the differences are not significant ( $p>0.05$ ).

### B. Method Comparison in Phase 2 (Indonesian)

On the Indonesian dataset (Table III), the four deep learning methods reached high and closely clustered accuracies (0.974–0.983), whereas Standalone DE declined sharply to 0.7755 with substantial variance. ANOVA was significant ( $F=7.19$ ;  $p=0.005$ ), yet Tukey HSD confirmed that the significant difference arises solely between Standalone DE and the other four methods; the proposed model does not differ significantly from LSTM, BiLSTM+GA, or CNN+PSO.

TABLE III  
PHASE 2 COMPARISON RESULTS (INDONESIAN)

| Method                     | Acc    | F1     | MER    | CR     |
|----------------------------|--------|--------|--------|--------|
| Proposed CNN-BiLSTM-Att+DE | 0.9742 | 0.9817 | 0.0258 | 0.0017 |
| Standalone LSTM            | 0.9830 | 0.9880 | 0.0170 | –      |
| CNN+PSO                    | 0.9780 | 0.9844 | 0.0220 | 0.0006 |
| BiLSTM+GA                  | 0.9802 | 0.9859 | 0.0198 | 0.0013 |
| Standalone DE              | 0.7755 | 0.8331 | 0.2245 | –      |

A notable finding is that the simplest model, Standalone LSTM, achieved the highest accuracy with a very short training time, indicating that the Phase 2 task is comparatively easy to learn. This relates to the lexical nature of LDA labels: any model can readily capture the same distinguishing patterns—a methodological limitation that warrants transparent reporting. The collapse of Standalone DE toward the majority class proportion demonstrates that DE without a hybrid architecture is inadequate, reinforcing the rationale for the proposed hybrid design and echoing findings from metaheuristic optimization studies on neural networks [22], [23].

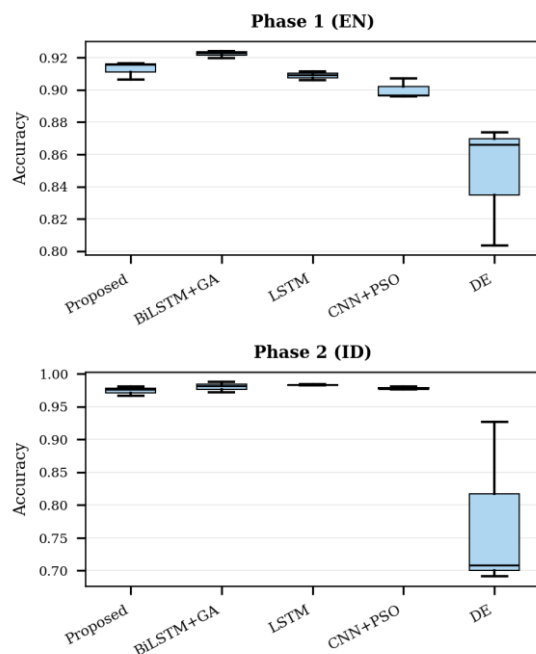


Fig. 3. Accuracy distribution of the five methods in Phase 1 (EN) and Phase 2 (ID).

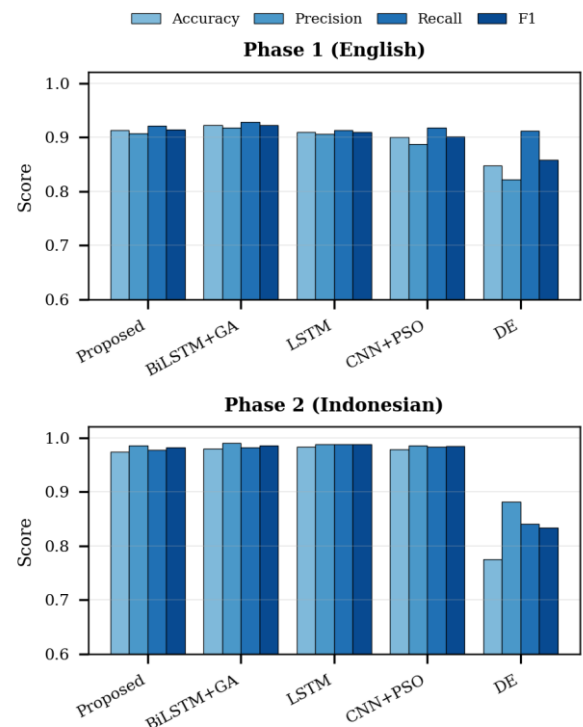


Fig. 4. Comparison of four metrics (Accuracy, Precision, Recall, F1) across the five methods in both phases.

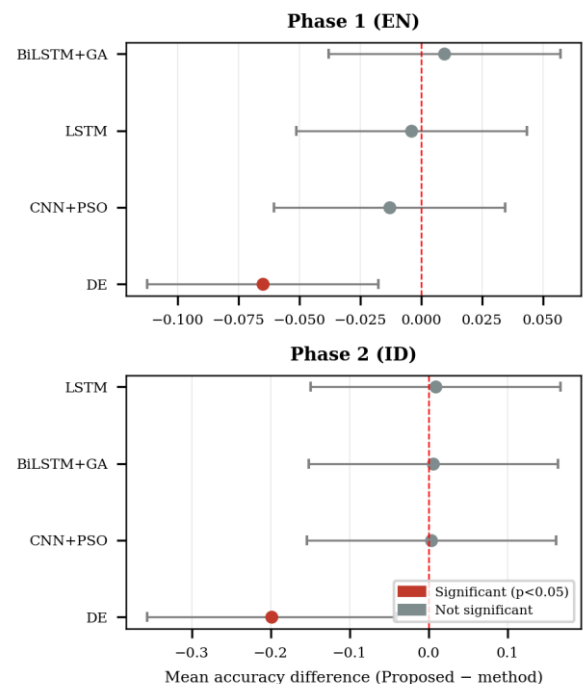


Fig. 5. Tukey HSD *post-hoc* test: mean accuracy difference between the proposed model and each baseline. Red markers indicate significant differences ( $p < 0.05$ ); grey markers indicate non-significant differences. The proposed model differs significantly only from Standalone DE.

### C. Cross-Dataset Analysis

Table IV compares accuracy across both phases. All deep learning methods achieved higher accuracy in Phase 2 (ID) than in Phase 1 (EN), producing negative differences ( $\Delta$ ). Standalone DE, by contrast, shows a positive  $\Delta$ . This pattern reflects dataset characteristics and the nature of LDA labeling rather than any intrinsic advantage of the models on a particular language; the high Indonesian accuracy is partly a consequence of label predictability.

TABLE IV  
CROSS-DATASET COMPARISON (MEAN ACCURACY)

| Method                     | Phase 1 (EN) | Phase 2 (ID) | $\Delta$ (EN-ID) |
|----------------------------|--------------|--------------|------------------|
| Proposed CNN-BiLSTM-Att+DE | 0.9128       | 0.9742       | -0.0614          |
| Standalone LSTM            | 0.9089       | 0.9830       | -0.0741          |
| CNN+PSO                    | 0.9000       | 0.9780       | -0.0780          |
| BiLSTM+GA                  | 0.9224       | 0.9802       | -0.0578          |
| Standalone DE              | 0.8478       | 0.7755       | +0.0723          |

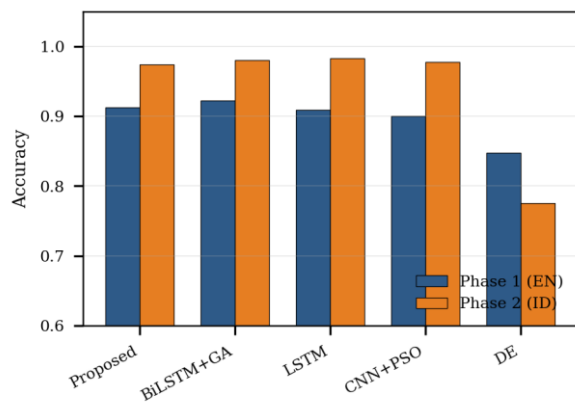


Fig. 6. Accuracy comparison of each method between Phase 1 (EN) and Phase 2 (ID).

### D. Cross-Domain Analysis

To assess generalization, the proposed model trained on product reviews (Phase 2) was evaluated on e-commerce application reviews from a different domain. Accuracy remained relatively stable (0.9774 in-domain versus 0.9753 cross-domain), but F1-Score fell sharply from 0.9839 to 0.8056, with precision at 0.7250. This decline indicates that the model does not generalize smoothly to a new domain and tends to mislabel genuine reviews as fake. This limited cross-domain generalization is consistent with the domain adaptation challenges reported in opinion spam detection [24]. Results are reported with the caveat of label imbalance in the target domain.

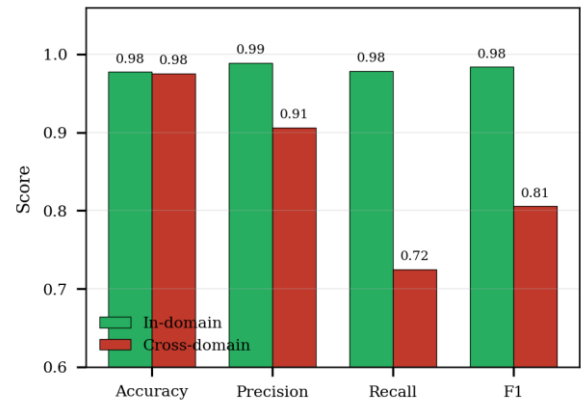


Fig. 7. In-domain versus cross-domain metrics for the proposed model. The pronounced drop in recall and F1 signals limited cross-domain generalization.

### E. Convergence Analysis of Optimization Algorithms

The Convergence Rate column in Tables II and III characterizes the behavior of the three metaheuristic algorithms. Standalone DE exhibits the highest convergence rate (0.0975 in Phase 1), yet this figure is misleading when read in isolation: rapid convergence occurs because the algorithm quickly becomes trapped in a solution that predicts the majority class, after which fitness ceases to improve. The proposed model, by contrast, shows a moderate convergence rate (0.0036) reflecting longer exploration of the search space before settling on a high-quality solution. This pattern underscores that convergence rate must always be interpreted alongside final quality metrics rather than as an indicator of superiority in its own right.

Comparison across algorithms shows that DE and GA display similar exploration characteristics, whereas PSO tends to converge faster but at slightly lower fitness values. This observation aligns with the review by Jiao et al. [22], which concluded that no evolutionary algorithm is universally superior; performance depends on the topography of the specific problem's search space.

### F. Error Analysis

Inspection of misclassified reviews revealed three dominant patterns. *First*, genuine reviews that are very short and generic, such as “bagus” (good) or “mantap sesuai deskripsi” (great, matches the description), are frequently flagged as CG because they lexically resemble machine-generated review patterns. *Second*, genuine reviews that are long and highly enthusiastic with many superlatives are similarly prone to misclassification, since superlatives constitute the principal marker the model learns from LDA labels. *Third*, code-mixed Indonesian-English reviews introduce representational ambiguity because the embeddings are trained on a predominantly Indonesian corpus.

All three patterns share a common root: LDA labels are formed from word distributions, so the model learns to separate writing styles rather than to distinguish genuine authenticity. This is the fundamental reason why Phase 2 accuracy is so high and, simultaneously, why cross-domain

performance deteriorates sharply. Reporting this finding openly is essential to avoid overstating what the impressive accuracy figures actually represent.

### *G. Discussion: Why the Proposed Model Is Not Uniformly Superior*

The results show that the proposed model is competitive yet does not consistently outperform every baseline. Three complementary explanations account for this. *First*, architectural complexity yields benefits only when the classification task is sufficiently difficult. In Phase 2, Standalone LSTM reached 0.9830 accuracy within a very short training time, signaling that the task is too easy for the additional capacity of CNN and attention to be exploited. *Second*, labels derived from LDA are lexical and internally consistent, so even simple models can separate them; the advantage of hybrid architectures emerges only on data with complex semantic patterns. *Third*, the DE evaluation budget equalized with PSO and GA for fairness constrains exploration of the higher-dimensional hyperparameter space of the proposed model.

The appropriate reading of these results is therefore narrower than the framing usually adopted in comparative studies. The proposed model is competitive with the strongest baselines; it is not demonstrably better than any of them. Tukey HSD finds no significant difference against Standalone LSTM, BiLSTM+GA, or CNN+PSO, and an absence of a significant difference is not evidence of superiority. The one difference the data do support is against Standalone DE, which lacks the hybrid architecture entirely. We therefore claim only that the architecture belongs to the top-performing group, and that the hybrid components matter more than the choice of optimizer.

A further boundary concerns what kinds of manipulation the present formulation can detect at all. The task as framed here is single-document and purely textual: a review is classified in isolation from its author, its timing, and the other reviews around it. Several forms of manipulation that dominate current practice fall outside that frame. Reviews produced by large language models are fluent, specific, and stylistically varied, which removes the lexical regularity that textual detectors rely on. Coordinated campaigns and review bombing are defined by the joint behaviour of many accounts over a short window rather than by any property of an individual text. Rating manipulation may leave the review body entirely genuine while the numerical score is inflated. Repeated or lightly edited submissions are visible only when documents are compared against each other.

Detecting these forms requires signals this study does not use: account histories, inter-arrival times of reviews, reviewer-product graphs, and rating-text consistency, of the kind Barbado et al. [24] found discriminative. A textual classifier remains useful as one component of such a system, but it should not be presented as a general solution to fake review detection. We regard the extension to behavioural and relational features as the single most valuable direction for subsequent work.

### *H. Practical Implications for MSMEs*

For MSMEs selling on e-commerce platforms, these results carry several applied implications. Models with high Precision are preferable in real deployment, because flagging genuine reviews as fake risks damaging the reputation of honest sellers. The drop in Precision to 0.7250 in cross-domain testing indicates that a detection system should not be transferred directly to a different product category without recalibration.

A realistic deployment positions the model as a first-stage screening tool that flags suspicious reviews for human review, rather than as an autonomous decision-maker. With an F1-Score above 0.97 within its matching domain, the model can meaningfully reduce the manual review workload while leaving final judgment to human assessors. Such a human-in-the-loop approach is also better suited to MSME contexts, where resources to absorb large-scale automated errors are typically unavailable.

Concretely, three deployment patterns fit the constraints of a micro or small seller. The first is reputation monitoring: the model scores incoming reviews on a seller's own listings and raises a weekly digest of the items whose review stream shifted most sharply, which is the signal a seller needs in order to contest a coordinated negative campaign with the platform. The second is triage of disputes: when a seller escalates a suspicious review, the model orders the queue so that the reviews most likely to be fabricated reach a human moderator first, which shortens resolution time without transferring the decision to the model. The third is pre-purchase advice on the buyer side, where flagged reviews are de-emphasized in an aggregate rating rather than removed.

All three patterns share a property that the measured precision makes necessary. A model that mislabels genuine reviews as fabricated imposes the cost of that error on an honest seller, who has no means of appeal against an automated verdict. None of these patterns should be implemented as autonomous removal, and the operating threshold should be tuned for precision rather than for balanced accuracy. We did not deploy or trial the system with actual sellers, so these remain design recommendations derived from the measured error profile rather than validated practice.

### *I. Limitations*

A limitation specific to the two-phase design concerns what the comparison between phases can support. The English and Indonesian conditions differ simultaneously in language, corpus size, class balance, review length, product domain, and, most consequentially, in how their labels were produced: the English labels accompany the dataset, whereas the Indonesian labels were induced by the procedure under test. Because these factors vary together, no difference between the phases can be attributed to language. We report the comparison as a cross-dataset observation and refrain from any claim about cross-lingual generalization, which would require a single dataset labeled by one procedure and available in both languages.

This study carries several limitations that warrant explicit acknowledgment. The most fundamental is label circularity: labels are generated by LDA from word distributions, and the model is then trained to predict those labels, so high accuracy partly reflects label predictability rather than a genuine capacity to detect fake reviews. The Cohen's Kappa of 0.939 does confirm that human annotators agree with these labels, but that agreement was formed on the same stylistic criteria.

A further limitation concerns feature representation: the model uses only textual features, without exploiting behavioral signals such as review timing patterns, account history, or user-graph connectivity, all of which proved informative in the study by Barbado et al. [24]. In addition, the datasets originate from three public corpora that do not necessarily represent the full population of Indonesian e-commerce reviews, and label imbalance in the target domain constrains the strength of cross-domain interpretation. Finally, annotation validation was performed on 200 samples; this quantity is statistically adequate but could be enlarged to increase confidence.

#### IV. CONCLUSION

This study successfully designed and implemented a hybrid CNN-BiLSTM-Attention architecture optimized by Differential Evolution (RQ1). For RQ2, the study reframes what automatic LDA labeling can establish. Agreement statistics measure whether annotators applied consistent criteria, not whether those criteria track authorship, so a high Kappa alone does not validate the labels. The revised procedure anchors the topic-to-class mapping to blind annotation and halts when that anchoring is not significantly better than chance. The cross-dataset study (RQ3) revealed EN-ID performance differences driven by dataset characteristics and labeling properties rather than by intrinsic language-specific model advantages.

Addressing the fourth research question, the comparison of five methods shows that the proposed model belongs to the highest-performing group alongside Standalone LSTM, BiLSTM+GA, and CNN+PSO, with no statistically significant differences among them, while Standalone DE without a hybrid architecture lags substantially and differs significantly. This finding confirms that metaheuristic optimization is not a substitute for an adequate architecture but a complement that performs best when paired with a suitable model structure.

The principal contribution of this work is a documented and testable labeling procedure for Indonesian OR/CG review data, together with a cross-dataset analysis situated in the MSME context. The main limitation is LDA label circularity, which renders the Phase 2 task comparatively easy and consequently understates the advantage of complex architectures.

Four directions are recommended for future research. *First*, replacing LDA-based labeling with a multi-feature scheme that combines user behavioral signals and review metadata, so that labels are no longer purely lexical. *Second*, expanding human annotation validation to a larger scale with annotation guidelines that define review authenticity

operationally rather than stylistically. *Third*, leveraging Indonesian pre-trained language models such as IndoBERT [18] as the representation layer to test whether hybrid architectures confer additional benefit on contextual representations. *Fourth*, evaluating performance on a subset of ambiguously labeled reviews to measure the advantage of complex architectures more fairly, since all models tend toward parity on easy data.

#### REFERENCES

- [1] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021. doi: 10.1109/ACCESS.2021.3075573.
- [2] R. Kumar, S. Mukherjee, and N. P. Rana, "Exploring latent characteristics of fake reviews and their intermediary role in persuading buying decisions," *Information Systems Frontiers*, vol. 25, pp. 1–16, 2023.
- [3] M. Q. Alnabhan and P. Branco, "Fake news detection using deep learning: A systematic literature review," *IEEE Access*, vol. 12, pp. 114435–114459, 2024. doi: 10.1109/ACCESS.2024.3435497.
- [4] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using LSTM-based deep learning models," *Frontiers in Big Data*, vol. 6, p. 1320800, 2024. doi: 10.3389/fdata.2023.1320800.
- [5] G. Bathla, P. Singh, R. K. Singh, E. Cambria, and R. Tiwari, "Intelligent fake reviews detection based on aspect extraction and analysis using deep learning," *Neural Computing and Applications*, vol. 34, no. 22, pp. 20213–20229, 2022. doi: 10.1007/s00521-022-07531-8.
- [6] W. Jian, J. Li, M. A. Akbar, A. U. Haq, S. Khan, R. M. Alotaibi, and S. A. Alajlan, "SA-Bi-LSTM: Self attention with bi-directional LSTM-based intelligent model for accurate fake news detection to ensured information integrity on social media platforms," *IEEE Access*, vol. 12, pp. 48436–48452, 2024. doi: 10.1109/ACCESS.2024.3383144.
- [7] S. P. Singh, N. Kumar, G. Kumar, and A. S. Shafeeq B. M., "A hybrid deep learning and differential evolution approach for accurate fake news detection," *Systems and Soft Computing*, vol. 7, p. 200365, 2025. doi: 10.1016/j.sasc.2025.200365.
- [8] N. Deshai and B. B. Rao, "Unmasking deception: A CNN and adaptive PSO approach to detecting fake online reviews," *Soft Computing*, vol. 27, no. 16, pp. 11357–11378, 2023. doi: 10.1007/s00500-023-08507-z.
- [9] A. Al Bataineh, W. Farhan, A. Alzubaidi, et al., "Advanced misinformation detection: A Bi-LSTM model optimized by genetic algorithms," *Electronics*, vol. 12, no. 15, p. 3250, 2023. doi: 10.3390/electronics12153250.
- [10] M. Baiocchi, G. Di Bari, A. Milani, and V. Poggioni, "Differential evolution for neural networks optimization," *Mathematics*, vol. 8, no. 1, p. 69, 2020. doi: 10.3390/math8010069.
- [11] K. Purifyregalia, K. Umam, N. C. H. Wibowo, and M. R. Handayani, "Detecting fake reviews in e-commerce: A case study on Shopee using Support Vector Machine and Random Forest," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, pp. 955–965, 2025. doi: 10.30871/jaic.v9i3.9514.
- [12] J. McAuley and J. Leskovec (mexwell), "Fake Reviews Dataset," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset>
- [13] R. Sutoyo, S. Achmad, A. Chowanda, E. W. Andangsari, and S. M. Isa, "PRDECT-ID: Indonesian product reviews dataset for emotions classification tasks," *Data in Brief*, vol. 44, p. 108554, 2022. doi: 10.1016/j.dib.2022.108554.
- [14] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012. doi: 10.11613/bm.2012.031.
- [15] D. Srikanth, K. Krishna Prasad, M. Kannan, and D. Kanchana, "Reliable social media framework: Fake news detection using

- modified feature attention based CNN-BiLSTM,” *International Journal of Machine Learning and Cybernetics*, vol. 16, pp. 2971–2996, 2025. doi: 10.1007/s13042-024-02431-w.
- [16] A. Romadhony, S. Al Faraby, R. Rismala, U. N. Wisesti, and A. Arifianto, “Sentiment analysis on a large Indonesian product review dataset,” *Journal of Information Systems Engineering and Business Intelligence (JISEBI)*, vol. 10, no. 1, pp. 167–178, 2024. doi: 10.20473/jisebi.10.1.167-178.
- [17] C.-H. Lin and U. Nuha, “Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy,” *Journal of Big Data*, vol. 10, no. 1, Art. no. 88, 2023. doi: 10.1186/s40537-023-00782-9.
- [18] 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. 28th International Conference on Computational Linguistics (COLING)*, pp. 757–770, 2020.
- [19] H. Murfi, T. Gowandi, G. Ardanewari, and S. Nurrohman, “BERT-based combination of convolutional and recurrent neural network for Indonesian sentiment analysis,” *Applied Soft Computing*, vol. 151, p. 111112, 2024.
- [20] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, “Stemming Indonesian: A confix-stripping approach,” *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 6, no. 4, Art. no. 13, pp. 1–33, 2007. doi: 10.1145/1316457.1316459.
- [21] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [22] R. Jiao, B. H. Nguyen, B. Xue, and M. Zhang, “A survey on evolutionary multiobjective feature selection in classification: Approaches, applications, and challenges,” *IEEE Transactions on Evolutionary Computation*, vol. 28, no. 4, pp. 1156–1176, 2024.
- [23] A. Sen, A. R. Mazumder, and U. Sen, “Differential evolution algorithm based hyper-parameters selection of transformer neural network model for load forecasting,” in *Proc. IEEE Symposium Series on Computational Intelligence (SSCI)*, 2023. doi: 10.1109/SSCI52147.2023.10371846.
- [24] R. Barbado, O. Araque, and C. A. Iglesias, “A framework for fake review detection in online consumer electronics retailers,” *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, 2019.
- [25] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, “Hybrid models for emotion classification and sentiment analysis in Indonesian language,” *Applied Computational Intelligence and Soft Computing*, vol. 2024, pp. 1–17, 2024.