

Fraud-Risk Classification in Indonesian MSME-Related Online Narratives: A Comparison of IndoBERT and TF-IDF-SVM

Bayu Priya Mukti ^{1*}

* Universitas Amikom Purwokerto
bayupriyaa@amikompurwokerto.ac.id ¹

Article Info

Article history:

Received 2026-08-02
Revised 2026-09-15
Accepted 2026-10-09

Keywords:

Digital Fraud,
IndoBert,
Indonesian NLP,
MSME,
Text Classification.

ABSTRACT

Digital fraud threatens micro, small, and medium enterprises (MSMEs), yet most automated detection studies remain centered on structured transactions, URLs, emails, or English-language benchmarks. This study evaluates fraud-risk narrative classification in 13,541 deduplicated Indonesian-language texts collected during a three-year window spanning 2023-2026 from Kompas, Detik, YouTube, and X. All retained documents were human-reviewed and assigned a binary discourse-level label and, for risk-positive records, one of seven fraud categories. A tuned TF-IDF linear support vector machine (SVM) was compared with IndoBERT using group-disjoint train-validation-test partitions, validation-only threshold selection, cost-sensitive learning, and 2,000 paired bootstrap resamples. On the untouched test set ($n=2,032$), IndoBERT achieved 0.9586 Macro-F1, 0.9636 accuracy, 0.9177 Matthews correlation coefficient, and 0.9771 PR-AUC. Its mean Macro-F1 improvement over SVM was 0.0149 with a 95% bootstrap confidence interval of 0.0045-0.0256, whereas the PR-AUC difference was small and uncertain (0.0022; 95% CI: -0.0044-0.0090). IndoBERT reduced false negatives from 69 to 50 and false positives from 31 to 24. Secondary analyses exposed uneven category difficulty, strong topic overlap around social engineering, and a length-sensitive trade-off: IndoBERT was better on short texts but worse than SVM on documents exceeding 512 words. The contribution is an auditable Indonesian MSME-related fraud benchmark that combines duplicate control, a strong classical baseline, calibrated decision thresholds, paired uncertainty, category-level diagnostics, and error analysis rather than proposing a new neural architecture.



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

I. INTRODUCTION

Digital fraud has become an operational risk for micro, small, and medium enterprises (MSMEs) as commercial activity increasingly depends on social media, marketplaces, mobile banking, and quick-response payment systems [1]-[3]. Fraud incidents are no longer limited to unauthorized card transactions. They may begin with impersonation, persuasive messages, malicious links, counterfeit QR codes, illegal lending offers, or fabricated investment opportunities. For small firms with limited cybersecurity resources, even a single incident can disrupt cash flow, expose credentials, damage customer trust, and create substantial recovery costs.

The research landscape does not yet represent this diversity equally. Reviews of digital payments and MSME adoption

describe rapid uptake and associated challenges, but generally focus on adoption, infrastructure, and payment-system risk rather than document-level fraud narratives [1]-[3]. Fraud-related text research remains fragmented by channel and attack type. Social-media studies emphasize automated or deceptive profiles [4]-[6]; recruitment-fraud detection operates in a narrow job-advertisement domain [7]; and phishing studies typically model emails, websites, URLs, or cybercrime artifacts [8]-[15]. Although fraud detection with natural-language processing is expanding [9], these settings differ substantially from heterogeneous Indonesian narratives that combine reportorial news, user comments, chronology, persuasion, warnings, and victim experience.

Transformer language models offer a plausible way to capture contextual signals that sparse lexical representations

may miss. Indonesian studies have demonstrated the value of contextual and hybrid deep representations for sentiment, multilabel, emotion, and review classification [16]-[20]. Foundational Indonesian resources, including IndoBERT, IndoBERTweet, and IndoNLU, provide the technical basis for such work [26]-[28]. Recent evidence also shows that the best representation strategy can depend on sample size and domain: frozen IndoBERT embeddings with XGBoost can rival full fine-tuning [32], while carefully designed TF-IDF pipelines remain strong for Indonesian user-generated text [33]. Nevertheless, evidence for digital-fraud narratives relevant to Indonesian MSMEs remains limited. More importantly, a transformer advantage is difficult to interpret when the comparator is weak, duplicates cross the data split, decision thresholds are optimized on the test set, or only point estimates are reported. Such practices create serious leakage and reproducibility risks [22], [23].

This study addresses both the domain gap and the evaluation gap. It asks four questions: (1) Does IndoBERT outperform a carefully tuned TF-IDF linear SVM for binary fraud-risk classification? (2) Is the observed improvement supported by paired uncertainty estimates? (3) Which fraud categories remain difficult under a secondary multiclass analysis? and (4) What do topic structure, lexical indicators, and document length reveal about model errors? The novelty lies not in a new architecture, but in an integrated benchmarking design: a deduplicated Indonesian MSME-oriented corpus, group-disjoint partitions, a tuned and threshold-calibrated classical baseline, an untouched test set, paired bootstrap comparison, seven-category diagnostics, and length-sensitive error analysis.

The resulting model is intended for prioritization and monitoring rather than autonomous legal or factual judgment. In practice, predicted risk can support content triage, analyst review, and targeted security education. Because the data include public narratives from different sources and writing styles, the analysis also provides a realistic test of how contextual and lexical models behave under heterogeneous Indonesian digital discourse.

II. MATERIALS AND METHODS

A. Data Sources and Label Structure

The raw corpus comprised 16,142 Indonesian-language documents collected during a three-year rolling window spanning 2023-2026 through source-specific API crawling: 3,556 records from Kompas, 3,500 from Detik, 5,709 from YouTube, and 3,377 from X. Kompas and Detik contributed online news reports, whereas the social-media portion consisted of YouTube comments and X posts or replies discussing QRIS-related digital fraud events. A document was eligible when it was written in Indonesian and contained at least one of the terms *qris*, *penipuan*, *digital*, or *umkm*.

Extremely short comments, exact or normalized duplicates, and records that were unrelated to the study context after content screening were excluded. User identifiers were not reported, and only aggregate results are presented.

The binary outcome was operationalized at the discourse level rather than from the mere occurrence of a fraud-related keyword. The fraud-risk class represented user-level comments, posts, or narratives describing exposure to fraud, attempted deception, victimization, or a concrete QRIS-related modus. The non-risk class primarily comprised reportorial or informational news texts. Such news records were assigned to the non-risk class even when they mentioned a fraud event or contained risk terminology, because third-person journalistic reporting was not treated as direct evidence of a target user-level fraud-risk narrative. Thus, non-risk in this study means absence of the target discourse-level account, not absence of fraud-related words. Risk-positive records were additionally assigned one primary category: social engineering, fake account, phishing, marketplace scam, fake QRIS, fraudulent investment, or illegal online loan. All 13,541 retained documents were independently checked by two Indonesian-language annotators, and disagreements were resolved through adjudication using a shared operational guide [24]. Binary agreement was 0.880 with Cohen's kappa of 0.822; category agreement was 0.810 with kappa of 0.800. The lower category agreement is substantively important because a single incident may combine impersonation, phishing, payment manipulation, and persuasion even though the dataset requires one primary category. All experiments, tables, and figures reported in this manuscript use this final human-reviewed corpus.

B. Data Audit, Deduplication, and Partitions

The eligibility screening and deduplication pipeline reduced 16,142 raw records to 13,541 retained documents. Extremely short texts, exact and normalized duplicates, and documents that failed the relevance criterion were removed before partitioning. In the final corpus, mean length was 1,443.1 characters or 199.8 words; median length was 222 words and the maximum was 1,702 words. No normalized duplicate group contained conflicting binary labels. Deduplication was completed before partitioning so that near-identical records could not cross data subsets.

Figure 1 introduces the final corpus. Panel (a) shows that 9,025 records (66.65%) were non-risk under the study's discourse-level definition and 4,516 (33.35%) were fraud risk. Panel (b) reveals a pronounced long tail: social engineering accounts for 2,703 risk-positive records, whereas illegal online loans contribute only 56. This imbalance motivates the use of Macro-F1 and category-specific diagnostics rather than accuracy alone.

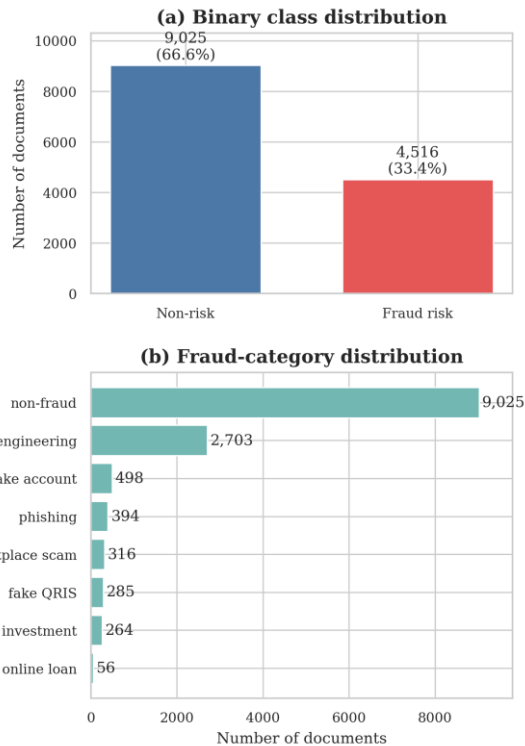


Figure 1. Binary class distribution and fraud-category composition after deduplication.

The deduplicated corpus was divided by normalized text group into training, validation, and test partitions using a fixed seed of 42. As detailed in Table I, the split retained nearly identical risk prevalence across subsets. Figure 2 provides the same composition visually and confirms that the validation and test sets are equal in size. A direct group-overlap audit found no shared text group across the three partitions.

TABLE I. CORPUS COMPOSITION AND GROUP-DISJOINT PARTITIONS

Partition	Non-risk	Risk	Total	% risk
Train	6,317	3,160	9,477	33.34
Validation	1,354	678	2,032	33.37
Test	1,354	678	2,032	33.37
Total	9,025	4,516	13,541	33.35

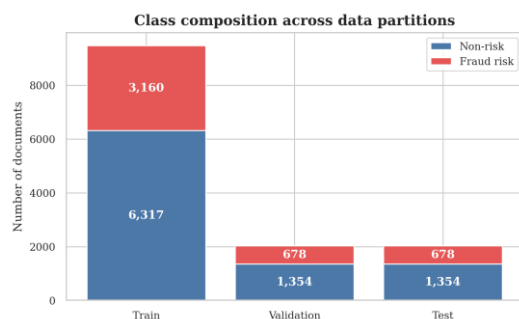


Figure 2. Class composition of the training, validation, and test partitions.

C. Tuned TF-IDF Linear Baseline

The classical baseline used word-level TF-IDF features with one- to three-word n-grams. A randomized search evaluated 15 configurations through five-fold cross-validation on the training partition. The best setting used $C=1.90475$, $\max_df=0.90$, a 30,000-feature cap, and class balancing. Before tuning, linear SVM reached 0.9449 validation Macro-F1, compared with 0.9377 for logistic regression and 0.5120 for a dummy classifier. The tuned SVM achieved 0.9450 cross-validated Macro-F1. Its decision threshold was then selected exclusively on the validation set, yielding a final threshold of 0.440.

D. IndoBERT Fine-tuning

The transformer model was indobenchmark/indobert-base-p1. The 95th percentile of tokenized length was 664 tokens; the maximum input length was limited to 256 tokens to fit the available GPU memory. Training used AdamW, a learning rate of 2×10^{-5} , weight decay of 0.01, a 0.10 warmup ratio, batch sizes of 16 for training and 32 for evaluation, two-step gradient accumulation, a linear scheduler, and mixed precision on an NVIDIA Tesla T4. Weighted cross-entropy used class weights [0.7501, 1.4995]. The best checkpoint was selected by validation Macro-F1 with early-stopping patience of two evaluations. The final validation-selected decision threshold was 0.905.

Table II consolidates the modeling protocol. The table is included to make the comparison auditable: both models share the same group-disjoint partitions and untouched test

set, but they differ in representation, optimization, and threshold. Threshold selection and checkpoint selection never used test labels.

TABLE II. MODELING AND SELECTION PROTOCOL

Component	TF-IDF + SVM	IndoBERT
Representation	1-3 word n-grams; 30k features	256-token contextual encoder
Imbalance	Balanced class weights	Weighted cross- entropy
Selection	15 configurations; 5- fold CV	Best validation Macro-F1
Training	C=1.90475	5 epochs; LR= 2×10^{-5}
Threshold	0.440 (validation)	0.905 (validation)
Test use	Final evaluation only	Final evaluation only

E. Evaluation and Secondary Analyses

Macro-F1 was the primary metric because it assigns equal importance to both classes and is less misleading than accuracy when class distributions are unequal [21]. Accuracy, balanced accuracy, risk-class precision, recall and F1, Matthews correlation coefficient (MCC), PR-AUC, ROC-AUC, and confusion matrices were also reported. Model differences were evaluated with 2,000 paired bootstrap resamples of the same test predictions. Percentile 95% confidence intervals were computed for each model and for the IndoBERT-minus-SVM difference. A difference was interpreted as supported only when its interval excluded zero.

Three secondary analyses were conducted. First, TF-IDF multinomial logistic regression classified the seven fraud categories within the risk-positive subset. Second, non-negative matrix factorization (NMF) summarized six latent lexical topics; NMF is appropriate for non-negative TF-IDF matrices and has been widely examined for short social-media text [29]-[31]. Recent work also explores large-language-model-based thematic analysis [25], but NMF was retained here because its term-topic structure is deterministic and directly inspectable. Third, global linear coefficients and record-level errors were analyzed. Length sensitivity compared records of 512 words or fewer with records exceeding 512 words. The cutoff was specified before inspecting model-specific error rates.

III. RESULTS AND DISCUSSION

A. Binary Test Performance

Table III reports the complete binary test results, while Figure 3 emphasizes the four metrics most relevant to imbalanced classification. IndoBERT achieved 0.9586

Macro-F1, 0.9636 accuracy, 0.9543 balanced accuracy, and 0.9177 MCC. The tuned SVM remained a strong comparator at 0.9439 Macro-F1 and 0.8886 MCC. The largest practical gain was in risk recall, which increased from 0.8982 to 0.9263 without sacrificing precision. Thus, the contextual model recovered more fraud-risk narratives while also slightly reducing false alarms.

TABLE III. BINARY PERFORMANCE ON THE UNTOUCHED TEST SET (N=2,032)

Metric	SVM	IndoBERT	Difference
Accuracy	0.9508	0.9636	+0.0128
Balanced accuracy	0.9377	0.9543	+0.0166
Risk precision	0.9516	0.9632	+0.0116
Risk recall	0.8982	0.9263	+0.0281
Risk F1	0.9241	0.9444	+0.0203
Macro-F1	0.9439	0.9586	+0.0147
MCC	0.8886	0.9177	+0.0291
PR-AUC	0.9749	0.9771	+0.0022
ROC-AUC	0.9817	0.9828	+0.0011

Test-set performance (95% bootstrap CIs where available)

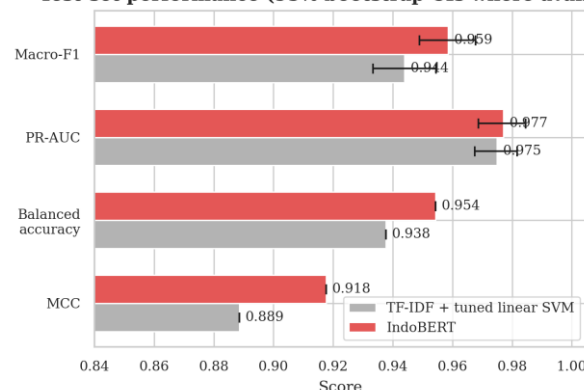


Figure 3. Comparison of test-set Macro-F1, PR-AUC, balanced accuracy, and MCC.

The source of this improvement is visible in Figure 4. The SVM produced 69 false negatives and 31 false positives; IndoBERT reduced these counts to 50 and 24, respectively. Across 2,032 test records, the contextual model therefore corrected 19 additional risk cases and seven additional non-risk cases at the aggregate confusion-matrix level. This balance matters operationally because a missed fraud-risk narrative can delay intervention, whereas excessive false positives increase the burden on analysts.

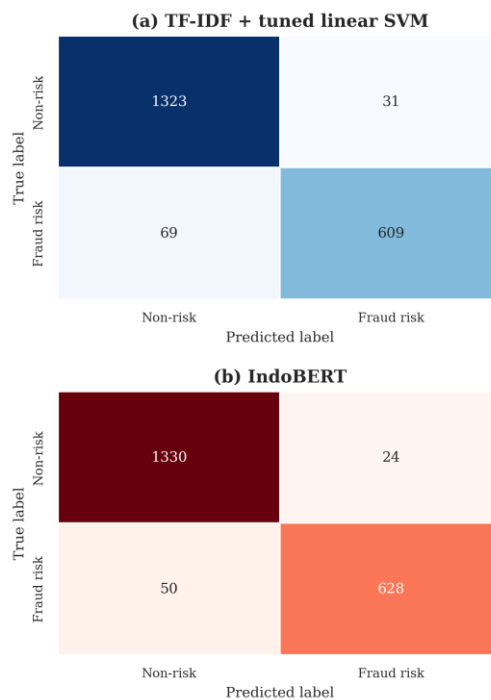


Figure 4. Confusion matrices for the tuned linear SVM and IndoBERT.

Figure 5 compares discrimination across all possible thresholds. The ROC curves are nearly superimposed, with AUC values of 0.982 and 0.983. The precision-recall curves show a similarly small separation, with average precision of 0.975 and 0.977. These curves explain why the fixed-

threshold gains in Macro-F1 and recall are more substantial than the differences in ranking metrics: both models rank most cases well, but IndoBERT provides a more favorable operating point after validation-based threshold selection.

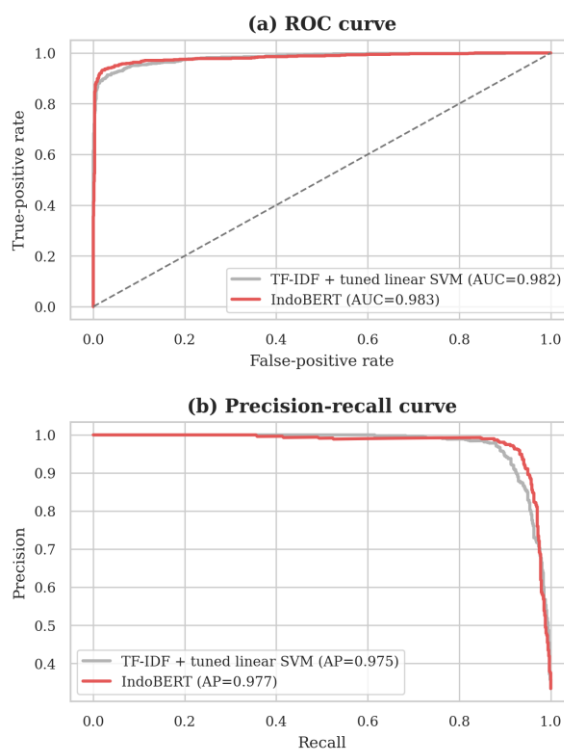


Figure 5. ROC and precision-recall curves for both binary classifiers.

B. Training Dynamics

Figure 6 shows the epoch-level learning dynamics. Training loss declined monotonically from 0.1867 to 0.0213, whereas validation loss reached its minimum near epoch 2 and then rose to 0.1875 by epoch 5. Validation Macro-F1 increased rapidly and stabilized around 0.96, while PR-AUC

remained near 0.983 after epoch 2. This divergence indicates increasing confidence on the training set without a corresponding ranking improvement on validation data. Checkpoint selection by validation Macro-F1 is therefore more defensible than reporting the final epoch automatically. Nevertheless, a single seed cannot quantify variability from parameter initialization or minibatch order.

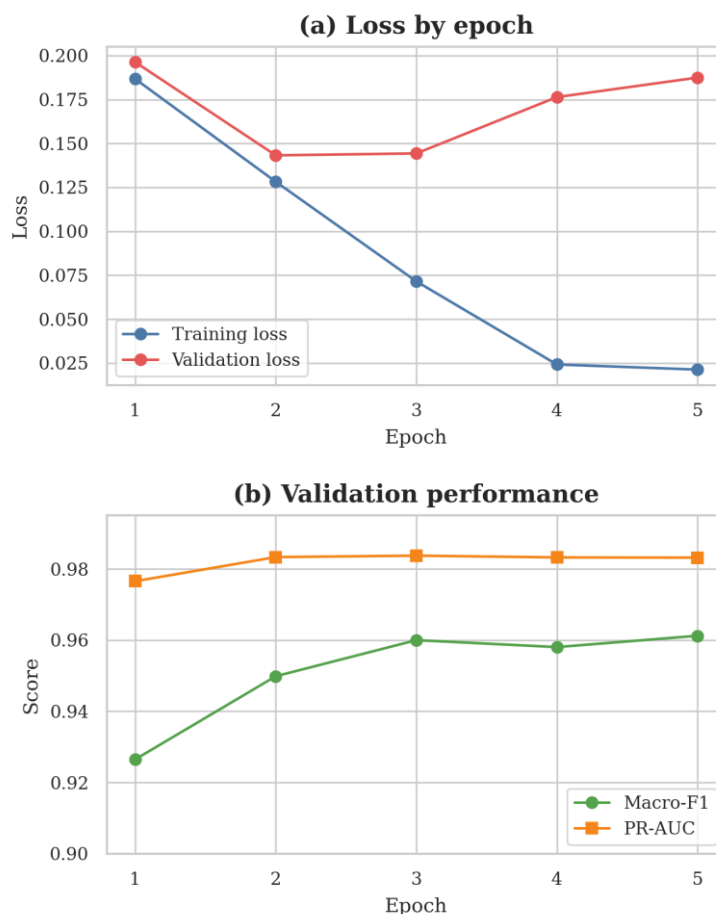


Figure 6. IndoBERT training loss, validation loss, Macro-F1, and PR-AUC by epoch.

C. Paired Bootstrap Uncertainty

Table IV separates score magnitude from uncertainty. The mean bootstrap Macro-F1 was 0.9438 for SVM and 0.9587 for IndoBERT. Their paired difference was 0.0149 with a 95% confidence interval of 0.0045-0.0256. Figure 7 shows the corresponding distribution: almost all resampled differences lie to the right of zero. The empirical probability of a non-positive difference was 0.0005, supporting the

conclusion that IndoBERT improves Macro-F1 on this test population.

The PR-AUC result leads to a different conclusion. Although IndoBERT's point estimate was higher, the paired difference of 0.0022 had a 95% interval from -0.0044 to 0.0090, and the empirical probability of a non-positive difference was 0.257. The study therefore does not claim a confirmed PR-AUC improvement. Reporting this distinction prevents a small point-estimate advantage from being presented as a robust gain.

TABLE IV. PAIRED BOOTSTRAP RESULTS FROM 2,000 RESAMPLES

Metric	SVM mean [95% CI]	IndoBERT mean [95% CI]	Paired difference [95% CI]
Macro-F1	0.9438 [0.9334, 0.9545]	0.9587 [0.9490, 0.9677]	0.0149 [0.0045, 0.0256]
PR-AUC	0.9749 [0.9675, 0.9816]	0.9771 [0.9687, 0.9846]	0.0022 [-0.0044, 0.0090]

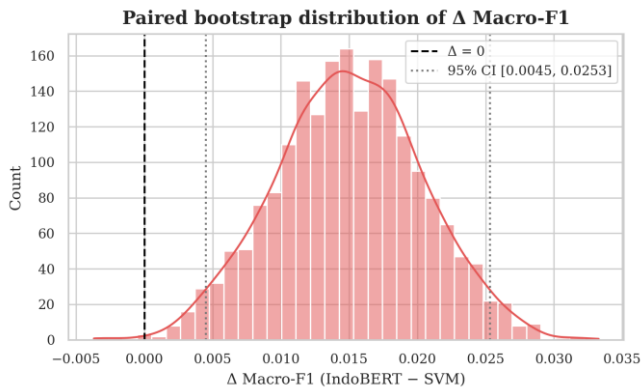


Figure 7. Paired bootstrap distribution of the IndoBERT-minus-SVM Macro-F1 difference.

D. Seven-Category Secondary Classification

The secondary multiclass baseline achieved 0.7950 accuracy and 0.7001 Macro-F1. Table V shows that fake QRIS and social engineering reached the highest F1 scores, 0.8764 and 0.8651, respectively. Fraudulent investment reached 0.7750, while phishing achieved high recall (0.8644) but lower precision (0.6456). Fake accounts and marketplace scams were more difficult, and the illegal-online-loan estimate was unstable because only eight cases were present in the secondary test partition.

TABLE V. SECONDARY MULTICLASS PERFORMANCE WITHIN RISK-POSITIVE RECORDS

Fraud category	Precision	Recall	F1	Support
Fake account	0.5976	0.6533	0.6242	75
Fraudulent investment	0.7750	0.7750	0.7750	40
Marketplace scam	0.5652	0.5532	0.5591	47
Phishing	0.6456	0.8644	0.7391	59
Illegal online loan	0.6000	0.3750	0.4615	8
Fake QRIS	0.8478	0.9070	0.8764	43
Social engineering	0.8947	0.8374	0.8651	406

Figure 8 clarifies where these errors occur. Seventeen fake-account cases were predicted as social engineering, while 27 social-engineering cases were predicted as fake accounts. Marketplace scams were confused with phishing nine times and with social engineering seven times. Four of eight illegal-online-loan cases were absorbed into social engineering. These patterns are consistent with the annotation

challenge: a message may impersonate a trusted party, direct the victim to a phishing channel, and request a transfer, but a single-label taxonomy forces one primary category.

		Fraud-category confusion matrix (secondary TF-IDF baseline)						
True category	fake acct.	49	1	1	5	2	0	17
	invest.	1	31	2	2	0	0	4
	market.	4	1	26	9	0	0	7
	phishing	1	0	1	51	0	1	5
	illegal loan	0	0	0	1	3	0	4
	fake QRIS	0	0	0	1	0	39	3
	social eng.	27	7	16	10	0	6	340
	Predicted category	fake acct.	invest.	market.	phishing	illegal loan	fake QRIS	social eng.

Figure 8. Confusion matrix for the seven-category TF-IDF secondary baseline.

E. Topic Structure and Lexical Evidence

Figure 9 presents the within-topic category proportions from the six-topic NMF solution. Social engineering dominates five topics, ranging from 0.61 in Topic 0 to 0.92 in Topic 1. Topic 2 is more heterogeneous: phishing contributes 0.24, fake accounts 0.20, marketplace scams 0.10, and social engineering 0.30. Fake QRIS reaches its largest share in Topic 3 at 0.12. The heatmap indicates that lexical topics do not map one-to-one onto fraud categories. Instead, social-engineering language cuts across incident types and channels, which helps explain both the dominance of this label and the off-diagonal multiclass errors.

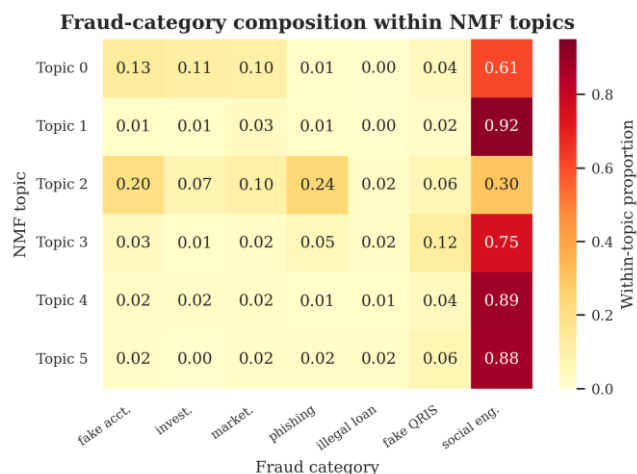


Figure 9. Fraud-category proportions within the six NMF topics.

The global linear explanation in Figure 10 shows why the classical baseline remained highly competitive. Its strongest risk indicators include penipuan (fraud), penipu (fraudster), ketipu (being scammed), modus, tipu, palsu (fake), ditipu, tertipu, OTP, scan QRIS, raib (lost), mengatasnamakan (impersonating), scam, and password. Many narratives therefore contain direct lexical evidence. At the same time, these coefficients reveal a failure mechanism: news reports and educational warnings may mention the same terms without describing a risk-positive event, while implicit incidents may describe persuasion and loss without using an explicit fraud word.

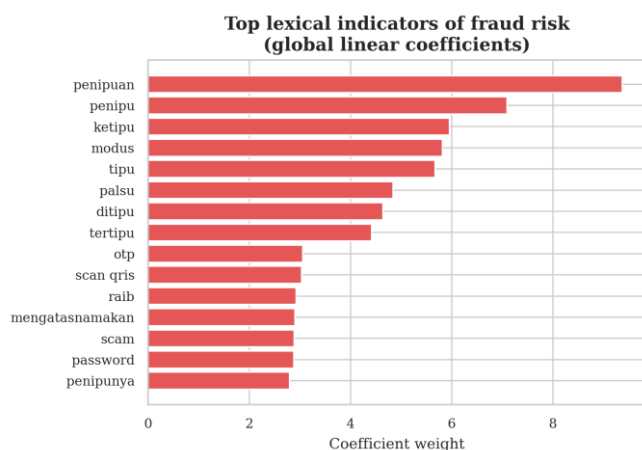


Figure 10. Fifteen strongest positive lexical coefficients in the global linear explanation.

F. Record-Level Error Analysis

The latest record-level analysis in Table VI goes beyond the aggregate confusion matrices. SVM made 100 errors and IndoBERT made 74. Forty-one cases were missed by both models; SVM alone was wrong on 59 cases, whereas IndoBERT alone was wrong on 33. The net reduction of 26 errors therefore reflects complementary behavior rather than

simple containment. IndoBERT reduced social-engineering errors from 42 to 26 and phishing errors from 12 to seven, but its fake-QRIS errors increased from 10 to 12. This category-specific reversal would be hidden by overall metrics.

TABLE VI. RECORD-LEVEL ERROR DECOMPOSITION ON THE BINARY TEST SET

Error diagnostic	SVM	IndoBERT
Total errors	100	74
False positives	31	24
False negatives	69	50
Social-engineering errors	42	26
Phishing errors	12	7
Fake-QRIS errors	10	12
Errors unique to model	59	33
Errors shared by both	41	41

Manual inspection of high-confidence disagreements revealed three recurring patterns. First, short comments sometimes contained fraud vocabulary but lacked enough context to establish a target user-level fraud account. Second, reportorial news articles contained fraud terminology, chronology, or warnings but remained informational under the operational labeling rule; lexical models could therefore interpret them as risk-positive. Third, informal spelling and compressed narratives could obscure the distinction between a personal incident, an attempted fraud, a warning, and a general reaction. These cases reinforce the need for periodic label-guideline review and a deployment option for uncertain cases.

G. Sensitivity to Document Length

Table VII and Figure 11 show that document length changes the relative ranking of the models. On 1,948 texts of 512 words or fewer, IndoBERT reduced the error rate from 0.0503 to 0.0364. On the 84 longer documents, however, SVM achieved a lower error rate (0.0238) than IndoBERT (0.0357). The long-document group also had lower fraud prevalence (22.62%) than the short-text group (33.83%), so truncation is not the only possible explanation. Nevertheless, the 256-token input limit makes it plausible that decisive information near the end of long articles was unavailable to IndoBERT, whereas TF-IDF retained evidence from the full document.

TABLE VII. ERROR RATE BY DOCUMENT-LENGTH GROUP

Length group	n	% risk	SVM error	IndoBERT error
≤512 words	1,948	33.83	0.0503	0.0364
>512 words	84	22.62	0.0238	0.0357

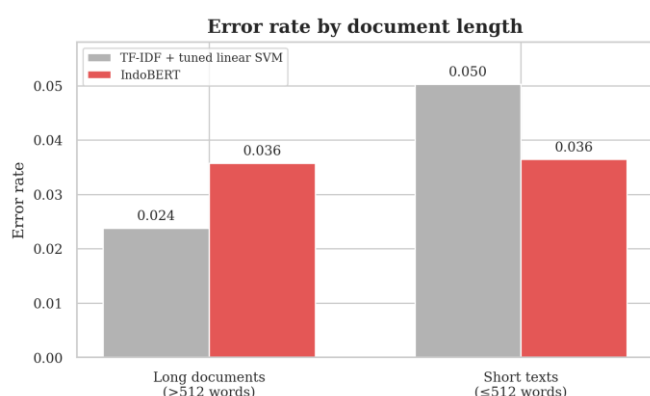


Figure 11. Error-rate comparison for short texts and long documents.

This trade-off suggests a practical hybrid policy. IndoBERT is preferable for short and contextually ambiguous records, while full-document linear features remain a credible low-cost fallback for long news articles. Future work should compare sliding-window aggregation, hierarchical encoders, and long-context transformers under source- and category-controlled evaluation. Any improvement should be tested prospectively rather than inferred from the present subgroup comparison.

H. Implications, Validity, and Limitations

The findings support IndoBERT as a triage model rather than a final decision maker. The system distinguishes target fraud-risk narratives from reportorial or informational texts; it does not estimate the prevalence of fraud among Indonesian MSMEs. Its validation-selected threshold produced high precision and recall under the observed test prevalence, but predictive values may change when deployed in a platform with a different risk base rate. Calibration monitoring, drift detection, and periodic threshold review are therefore necessary. The seven-category output can support routing-for example, phishing cases toward credential response and fake-QRIS cases toward merchant verification-but rare categories should be allowed to enter a review state rather than being forced into a confident automated decision.

Internal validity was strengthened by deduplication before splitting, group-disjoint partitions, a tuned classical comparator, validation-only threshold selection, an untouched test set, and paired bootstrap analysis. These choices directly address the leakage and reproducibility concerns raised in [22], [23]. However, the same validation partition was used for checkpoint selection and threshold selection, which may introduce mild optimism. A nested or second-stage validation design would provide a stricter estimate.

Eight limitations remain. First, although the source totals and 2023-2026 collection window are known, document-level timestamps and location metadata were not sufficiently complete for temporal or geographic validation. Second, query-driven API sampling can overrepresent explicit, highly

visible, or highly engaged QRIS-related fraud discussions and underrepresent private, unreported, or implicitly described incidents. Third, online news and social-media comments do not represent the full range of fraud experienced by Indonesian MSMEs. Fourth, the operational rule places reportorial news predominantly in the non-risk class, creating possible source-label coupling; part of the observed discrimination may therefore reflect discourse or platform style. Fifth, the fraud-category distribution is highly imbalanced and category boundaries remain uncertain despite human review and adjudication. Sixth, only one random seed and one fixed partition were evaluated. Seventh, the 256-token limit constrains long-document coverage. Eighth, the test set is in-domain; no prospective evaluation was conducted on later periods, new platforms, or independent MSME institutions. Future work should add repeated group splits, multi-seed fine-tuning, probability calibration, temporal and source-based holdouts, multilabel annotation, and external validation.

IV. CONCLUSION

A comparison on 13,541 deduplicated Indonesian digital narratives showed that IndoBERT achieved 0.9586 Macro-F1 and 0.9177 MCC, outperforming a tuned TF-IDF linear SVM. The paired Macro-F1 improvement was supported by a 95% bootstrap interval that excluded zero, and the gain was accompanied by reductions in both false negatives and false positives. In contrast, the small PR-AUC improvement was not statistically supported because its interval crossed zero.

The secondary analyses qualify this overall result. Category performance was uneven, social engineering overlapped strongly with other topics and labels, and IndoBERT's advantage reversed on documents longer than 512 words. The most defensible contribution is therefore not a universal superiority claim, but an auditable benchmark showing where contextual modeling helps, where a linear model remains competitive, and how uncertainty and subgroup analysis alter interpretation. The reported scores apply to MSME-related online narratives collected through the specified news and social-media sources; they should not be interpreted as fraud-prevalence estimates or as performance on internal MSME transaction data. With external validation and operational monitoring, the framework can support fraud-content prioritization and security education for Indonesian MSMEs.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the two data annotators for their valuable contributions to the review data annotation. We also thank Universitas Amikom Purwokerto for its support and the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemendikristek) for funding this research through the 2026 Penelitian Dosen Pemula (PDP) grant program.

CONFLICT OF INTEREST

The author declares no conflict of interest that could have influenced the design, analysis, or reporting of this study.

REFERENCES

- [1] K. Khando, M. S. Islam, and S. Gao, "The emerging technologies of digital payments and associated challenges: A systematic literature review," *Future Internet*, vol. 15, no. 1, Art. no. 21, 2023, doi: 10.3390/fi15010021.
- [2] Saripudin, R. Yuniarti, and D. Ernawati, "Exploring the factors influencing the adoption of QRIS as a digital payment in Indonesia," *The Journal of Management Theory and Practice*, vol. 4, no. 1, pp. 53–59, 2023, doi: 10.37231/jmtp.2023.4.1.322.
- [3] A. B. P. Irianto and P. Chanvarasuth, "Drivers and barriers of mobile payment adoption among MSMEs: Insights from Indonesia," *Journal of Risk and Financial Management*, vol. 18, no. 5, Art. no. 251, 2025, doi: 10.3390/jrfm18050251.
- [4] M. Aljabri, R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, and D. M. Alomari, "Machine learning-based social media bot detection: A comprehensive literature review," *Social Network Analysis and Mining*, vol. 13, Art. no. 20, 2023, doi: 10.1007/s13278-022-01020-5.
- [5] B. G. Bokolo and Q. Liu, "Advanced algorithmic approaches for scam profile detection on Instagram," *Electronics*, vol. 13, no. 8, Art. no. 1571, 2024, doi: 10.3390/electronics13081571.
- [6] B. Goyal, N. S. Gill, and P. Gulia, "Securing social spaces: Machine learning techniques for fake profile detection on Instagram," *Social Network Analysis and Mining*, vol. 14, Art. no. 231, 2024, doi: 10.1007/s13278-024-01399-3.
- [7] M. Naudé, K. J. Adebayo, and R. Nanda, "A machine learning approach to detecting fraudulent job types," *AI & Society*, vol. 38, pp. 1013–1024, 2023, doi: 10.1007/s00146-022-01469-0.
- [8] M. Schmitt and I. Flechais, "Digital deception: Generative artificial intelligence in social engineering and phishing," *Artificial Intelligence Review*, vol. 57, Art. no. 324, 2024, doi: 10.1007/s10462-024-10973-2.
- [9] P. Boulrieris, J. Pavlopoulos, A. Xenos, and V. Vassalos, "Fraud detection with natural language processing," *Machine Learning*, vol. 113, pp. 5087–5108, 2024, doi: 10.1007/s10994-023-06354-5.
- [10] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University–Computer and Information Sciences*, vol. 35, pp. 590–611, 2023, doi: 10.1016/j.jksuci.2023.01.004.
- [11] K. Thakur, M. L. Ali, M. A. Obaidat, and A. Kamruzzaman, "A systematic review on deep-learning-based phishing email detection," *Electronics*, vol. 12, no. 21, Art. no. 4545, 2023, doi: 10.3390/electronics12214545.
- [12] S. Atawneh and H. Aljehani, "Phishing email detection model using deep learning," *Electronics*, vol. 12, no. 20, Art. no. 4261, 2023, doi: 10.3390/electronics12204261.
- [13] N. Altwaijry, I. Al-Turaiki, R. Alotaibi, and F. Alakeel, "Advancing phishing email detection: A comparative study of deep learning models," *Sensors*, vol. 24, no. 7, Art. no. 2077, 2024, doi: 10.3390/s24072077.
- [14] R. Meléndez, M. Ptaszynski, and F. Masui, "Comparative investigation of traditional machine-learning models and transformer models for phishing email detection," *Electronics*, vol. 13, no. 24, Art. no. 4877, 2024, doi: 10.3390/electronics13244877.
- [15] F. S. Alsubaiei, A. A. Almazroi, and N. Ayub, "Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics," *IEEE Access*, vol. 12, pp. 8373–8389, 2024, doi: 10.1109/ACCESS.2024.3351946.
- [16] C.-H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *Journal of Big Data*, vol. 10, Art. no. 88, 2023, doi: 10.1186/s40537-023-00782-9.
- [17] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardanawari, and S. Nurrohman, "BERT-based combination of convolutional and recurrent neural network for Indonesian sentiment analysis," *Applied Soft Computing*, vol. 151, Art. no. 111112, 2024, doi: 10.1016/j.asoc.2023.111112.
- [18] G. Z. Nabillah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.
- [19] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of Indonesian tweets using large scale transfer learning via IndoBERT," *Social Network Analysis and Mining*, vol. 15, Art. no. 22, 2025, doi: 10.1007/s13278-025-01439-6.
- [20] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, 2024, doi: 10.11591/eei.v13i5.8032.
- [21] A. E. Blanchard et al., "A keyword-enhanced approach to handle class imbalance in clinical text classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2796–2803, 2022, doi: 10.1109/JBHI.2022.3141976.
- [22] A. Apicella, F. Isgro, and R. Prevete, "Don't push the button! Exploring data leakage risks in machine learning and transfer learning," *Artificial Intelligence Review*, vol. 58, Art. no. 339, 2025, doi: 10.1007/s10462-025-11326-3.
- [23] R. Joeres, D. B. Blumenthal, and O. V. Kalinina, "Data splitting to avoid information leakage with DataSAIL," *Nature Communications*, vol. 16, Art. no. 3337, 2025, doi: 10.1038/s41467-025-58606-8.
- [24] D. Braun, "I beg to differ: How disagreement is handled in the annotation of legal machine learning data sets," *Artificial Intelligence and Law*, vol. 32, no. 3, pp. 839–862, 2024, doi: 10.1007/s10506-023-09369-4.
- [25] C. Wen, P. Clough, R. Paton, and R. Middleton, "Leveraging large language models for thematic analysis: A case study in the charity sector," *AI & Society*, vol. 41, pp. 731–748, 2026, doi: 10.1007/s00146-025-02487-4.
- [26] 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*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.
- [27] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," in *Proc. EMNLP*, pp. 10660–10668, 2021, doi: 10.18653/v1/2021.emnlp-main.833.
- [28] B. Willie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. AACL-JCNLP*, pp. 843–857, 2020, doi: 10.18653/v1/2020.aacl-main.85.
- [29] R. Egger and J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts," *Frontiers in Sociology*, vol. 7, Art. no. 886498, 2022, doi: 10.3389/fsoc.2022.886498.
- [30] S. Athukorala and W. Mohotti, "An effective short-text topic modelling with neighbourhood assistance-driven NMF in Twitter," *Social Network Analysis and Mining*, vol. 12, Art. no. 89, 2022, doi: 10.1007/s13278-022-00898-5.
- [31] C. Doogan, W. Buntine, and H. Linger, "A systematic review of the use of topic models for short text social media analysis," *Artificial Intelligence Review*, vol. 56, pp. 14223–14255, 2023, doi: 10.1007/s10462-023-10471-x.
- [32] N. A. Putri, R. E. Saputro, D. I. S. Saputra, and B. P. Mukti, "IndoBERT embeddings vs. fine-tuning: An optimized approach for sentiment analysis of Indonesian public service reviews," in *Proc. 2026 International Conference on Current Research in Artificial Intelligence and Data Science (ICCRAIDS)*, Banyumas, Indonesia, Apr. 10–11, 2026, doi: 10.1109/ICCRAIDS67816.2026.11519745.
- [33] Z. Rifa'i and B. P. Mukti, "Weakly supervised sentiment analysis of Indonesian rural tourism reviews: A TF-IDF baseline for Melung Tourism Village," *Edu Komputika Journal*, vol. 12, no. 1, pp. 48–60, Aug. 2025, doi: 10.15294/edukom.v12i1.31893.