

Context-Aware Detection of Content Evidencing Illegal or Predatory Fintech Lending Practices on Indonesian Digital Platforms Using Transformer-Based Language Models

Uci Suriani¹, Netti Herawati¹, Ilsa Palingga Ninditama²

^{1,2}Multimedia Engineering Technology Study Program, Politeknik Prasetiya Mandiri, Indonesia

Received:

May 4, 2026

Revised:

September 5, 2026

Accepted:

September 17, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Uci Suriani

Email*:

uci.dosen@prasetiyamandiri.ac.id

DOI:

10.63158/journalisi.v8i5.1903

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. Illegal and predatory fintech lending practices generate digital content that may be difficult to identify using keyword-based screening because harmful practices can be expressed through formal, colloquial, and context-dependent language. This study benchmarks transformer-based language models for detecting content evidencing illegal or predatory fintech lending practices and examines whether Indonesian-specific pretraining provides an advantage over multilingual pretraining. A corpus of 10,000 Indonesian texts was collected from news, complaint platforms, regulatory reports, forums, consumer advocacy sources, and social media. The texts were annotated using six predefined indicators of harmful lending practices and normalised using a domain-specific slang dictionary and Sastrawi stemming. Multilingual BERT, IndoBERT, and IndoBERTtweet were fine-tuned using weighted cross-entropy and compared with a TF-IDF logistic regression baseline. Multilingual BERT achieved the highest macro F1-score (0.9616), followed by IndoBERT (0.9582) and IndoBERTtweet (0.9449), compared with 0.8974 for the baseline. Multilingual BERT missed only 9 of 413 harmful-content instances, reducing missed cases by 77.5% relative to the baseline. Indonesian-specific pretraining showed no clear advantage on the predominantly formal corpus, while stemming may reduce its benefit. The findings support transformer-based screening as a research benchmark for regulatory content monitoring, while human review remains necessary for enforcement decisions.

Keywords: context-aware text classification, illegal fintech lending, IndoBERT, natural language processing, transformer fine-tuning

1. INTRODUCTION

Peer-to-peer lending platforms operating without a licence from the Indonesian Financial Services Authority remain a persistent consumer protection problem in Indonesia. During 2025, the national task force blocked more than 1,500 unlicensed online lending operations, while the Indonesia Anti-Scam Centre recorded more than 320,000 related complaints [1]. However, enforcement actions against platforms do not eliminate the digital content associated with illegal lending activities. Promotional messages, borrower complaints, testimonies describing harmful practices, and other narratives may continue to circulate through news portals, online forums, and social media [2]. The resulting harm may include financial loss, intimidation, unauthorised access to borrowers' contact information, and psychological pressure during debt collection [3]. These conditions create a need for automated screening that can identify textual evidence of harmful or predatory lending practices across heterogeneous Indonesian digital sources.

In this study, the classification target is not all content that mentions illegal fintech lending. The positive class represents content that promotes or provides textual evidence of harmful or predatory lending practices, including coercive collection, unauthorised access to borrower data, debt entrapment, explicit borrower harm, fraud or impersonation, and psychological impact. Regulatory announcements, news reports, or neutral discussions that merely mention illegal lenders without evidence of such practices are treated as class 0.

Current operational screening commonly relies on blocklists and keyword-based filters, but these approaches are vulnerable to the linguistic variability of Indonesian digital communication. Users may combine formal and colloquial registers, abbreviate words inconsistently, or describe coercive collection practices without explicitly naming a lending platform. Expressions such as *galbay*, referring to *gagal bayar*, and *DC*, referring to *debt collector*, illustrate how informal terminology can vary across sources. More importantly, keyword matching cannot reliably distinguish between regulatory or neutral content that mentions illegal lending and content that promotes or provides evidence of harmful lending practices. Both may contain the same lending-related vocabulary, while the discriminative information often lies in the relationship between words and the meaning conveyed by the surrounding sentence [4]. This limitation motivates the use of

contextual language representations that can model relationships between tokens rather than treating words as independent lexical features.

Previous text-mining studies have addressed several forms of illicit or problematic Indonesian digital content, but their approaches and target domains remain different from the problem considered in this study. Suryono et al. [2] investigated fintech peer-to-peer lending issues in Indonesia through qualitative analysis of online news rather than automated classification. Perdana et al. [5] detected online gambling promotions on Indonesian Twitter using TF-IDF with Random Forest, reporting 0.966 precision and 0.958 recall on 6,038 tweets. In the broader fintech fraud literature, Stojanovic et al. [6] showed that many deployed fraud-detection systems rely primarily on transaction records rather than textual evidence. Indonesian-language research has also progressed in other content-moderation domains, including toxic-language detection using semantic expansion with bidirectional GRU [7], multi-label toxic speech classification [8], fake-news detection [9], and prediction tasks in education and health [10], [11]. Although these studies demonstrate the value of text mining and machine learning for Indonesian-language classification, they do not directly address context-aware detection of content evidencing illegal or predatory fintech lending practices across heterogeneous digital platforms.

Transformer-based language models provide a stronger alternative because they represent words in relation to their surrounding context. BERT learns bidirectional contextual representations through masked language modelling [12] based on the self-attention architecture introduced by Vaswani et al. [13]. Its multilingual version supports 104 languages, including Indonesian. Several Indonesian-specific alternatives have subsequently been developed. IndoBERT was pretrained on large Indonesian corpora, including Indo4B, and released with the IndoNLU benchmark [14], while another IndoBERT variant was trained using IndoLEM [15]. IndoBERTtweet further adapts Indonesian BERT to Twitter through domain-specific vocabulary initialization [16]. Previous comparative studies have often reported advantages for Indonesian-specific models. Nabilah et al. [17] reported an F1-score of 0.890, while Nugraha et al. [18] obtained 98% accuracy with IndoBERT compared with 96% for multilingual BERT on gambling-related comments. However, evidence also suggests that large-scale multilingual pretraining can narrow the performance difference between multilingual and monolingual models [19]. In addition,

Khairani et al. [20] found that removing stopwords and applying stemming reduced classification accuracy for both IndoBERT and IndoBERTweet, suggesting that the benefit of Indonesian-specific pretraining may depend on the preprocessing strategy and the characteristics of the input text.

These findings reveal two specific research gaps. First, based on the literature reviewed in this study, pretrained language models have not been directly evaluated for detecting textual evidence of illegal or predatory fintech lending practices across heterogeneous Indonesian digital platforms. Existing work in the Indonesian fintech-lending domain has primarily relied on qualitative analysis or lexical approaches, while transformer-based contextual classification has been more extensively examined in other domains such as gambling, toxic language, and fake news [2], [5], [7-11]. Second, the reported advantage of Indonesian-specific monolingual pretraining has mainly been demonstrated on informal or social-media-oriented benchmarks, leaving its effectiveness on a corpus dominated by formal news, regulatory reports, and complaint narratives insufficiently examined [17-19]. This distinction is important because pretrained vocabulary and domain-specific representations may interact differently with formal and colloquial Indonesian and with the preprocessing operations applied before fine-tuning [20].

Therefore, this research intends to evaluate whether contextual transformer representations provide a measurable advantage over conventional lexical classification and whether Indonesian-specific pretraining remains advantageous when the target corpus contains heterogeneous but predominantly formal Indonesian text. The study constructs an annotated corpus of 10,000 Indonesian texts collected from multiple digital source categories and compares multilingual BERT, IndoBERT, and IndoBERTweet under identical fine-tuning conditions against a TF-IDF logistic regression baseline. The objectives of this research are threefold: (1) to evaluate the contribution of contextual transformer representations to detecting textual evidence of illegal or predatory fintech lending practices compared with a lexical baseline; (2) to determine whether Indonesian-specific monolingual pretraining provides a measurable performance advantage over multilingual BERT in this domain; and (3) to characterise the precision-recall and error trade-offs relevant to automated regulatory content screening.

2. METHODS

2.1 Research design and the meaning of context

The study employed an experimental design consisting of six sequential stages: data acquisition, data annotation, text preprocessing, data splitting, model development, and model evaluation, as illustrated in Figure 1. All experiments were conducted using Python 3, PyTorch, and Hugging Face Transformers on a single NVIDIA A100 GPU. The same data partitions and evaluation procedures were applied to all models to ensure that performance differences primarily reflected differences in pretrained representations rather than variations in the training data or experimental conditions.

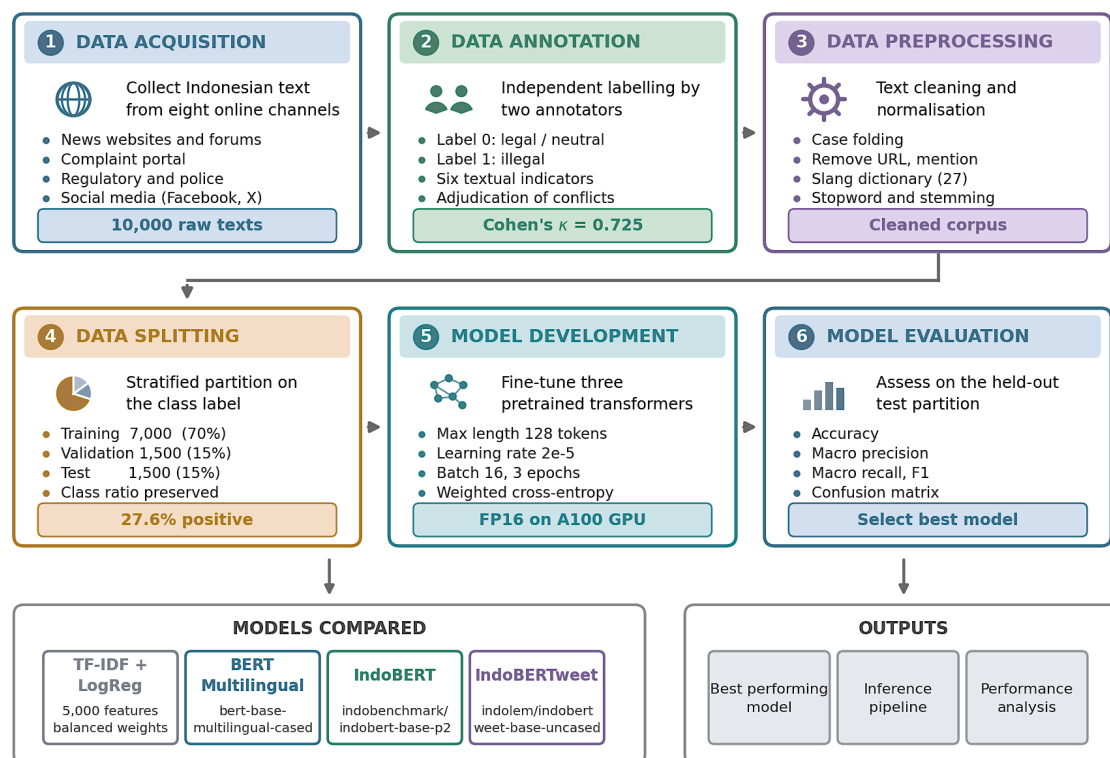


Figure 1. Research pipeline from data acquisition to model evaluation

Figure 1 illustrates the overall research pipeline, beginning with the acquisition of Indonesian texts from multiple digital sources and continuing through annotation, preprocessing, stratified data partitioning, model fine-tuning, and final evaluation. In operational terms, the annotated corpus is normalised through a fixed preprocessing sequence and then partitioned using stratified sampling. Three transformer models are subsequently fine-tuned under identical training configurations, with class imbalance

addressed through weighted cross-entropy. Model selection is based exclusively on validation macro F1, while the held-out test partition is used only for final evaluation. This procedure ensures a consistent comparison between the contextual transformer models and the TF-IDF baseline.

In this study, context-aware detection refers to classification based on relationships between tokens and their surrounding textual context rather than on the independent presence of specific keywords. For example, repeated contact directed at a borrower may indicate coercive collection even when no single term explicitly identifies the behaviour. The six annotation indicators in Table 2 therefore require interpretation of relationships among words and expressions rather than simple keyword matching.

The three transformer models represent text through self-attention, allowing the representation of each token to incorporate information from other tokens in the same sequence. The self-attention operation is defined as Equation 1.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q , K , and V denote the query, key, and value matrices, respectively, and d_k represents the dimensionality of the key vectors. Through this mechanism, the representation of a token is conditioned on the other tokens in the sequence, allowing the models to distinguish expressions with similar vocabulary but different contextual meanings. The TF-IDF baseline, in contrast, provides a context-free control because it represents the same texts primarily through term-frequency and inverse-document-frequency statistics. The performance difference between these two model families therefore provides an empirical indication of the contribution of contextual representation to the detection task.

2.2 Corpus construction and annotation

Indonesian texts concerning online lending were collected from eight categories of public sources published between 2022 and 2025. The sources included news websites, complaint platforms, regulatory and law-enforcement reports, consumer advocacy organisations, published research corpora, online forums, Facebook, and Twitter/X. After

removing duplicate records and texts containing fewer than two tokens, 10,000 instances remained. Table 1 presents the composition of the corpus by source channel.

Table 1. Composition of the corpus by source channel

No	Source channel	Instances	Share (%)
1	News website	3,278	32.78
2	Complaint portal (MediaKonsumen)	2,086	20.86
3	Official report (regulator, police, task force)	1,448	14.48
4	Consumer advocacy organisation	1,244	12.44
5	Other and published research corpora	763	7.63
6	Online forum	515	5.15
7	Social media (Facebook)	341	3.41
8	Social media (Twitter/X)	325	3.25
	Total	10,000	100.00

Each text was assigned a binary label based on predefined annotation criteria. Class 1 represents content that promotes or provides textual evidence of harmful or predatory lending practices, while class 0 represents regulatory, news, or other neutral content that mentions illegal lending without providing evidence of such practices. Six indicators were defined before annotation: coercive collection, unauthorised data access, debt entrapment, explicit borrower harm, fraud or impersonation, and psychological impact. A text was assigned to class 1 when at least one indicator was explicitly evidenced in the text rather than inferred solely from external information. Table 2 presents the indicators and corresponding textual evidence.

Table 2. Annotation indicators used to assign the illegal class

No	Indicator	Textual evidence sought
1	Coercive collection	Threats, abusive language, or repeated unsolicited contact with a borrower
2	Unauthorised data access	Contact lists, emergency contacts, or gallery data taken beyond stated consent

No	Indicator	Textual evidence sought
3	Debt entrapment	Cyclical borrowing or charges the borrower cannot settle
4	Explicit borrower harm	Stated financial loss, victimisation, or a complaint filed against a lender
5	Fraud or impersonation	Use of a licensed institution name, fabricated approval, or unrequested disbursement
6	Psychological impact	Reported fear or distress attributable to lending or collection

A graduate student and a lecturer independently annotated a subset of 1,196 texts drawn from the validation and test data. Inter-annotator agreement was measured using Cohen's kappa [21], as defined in Equation 2, where p_o represents observed agreement and p_e represents chance agreement.

$$AK = \frac{p_o - p_e}{1 - p_e} \quad (2)$$

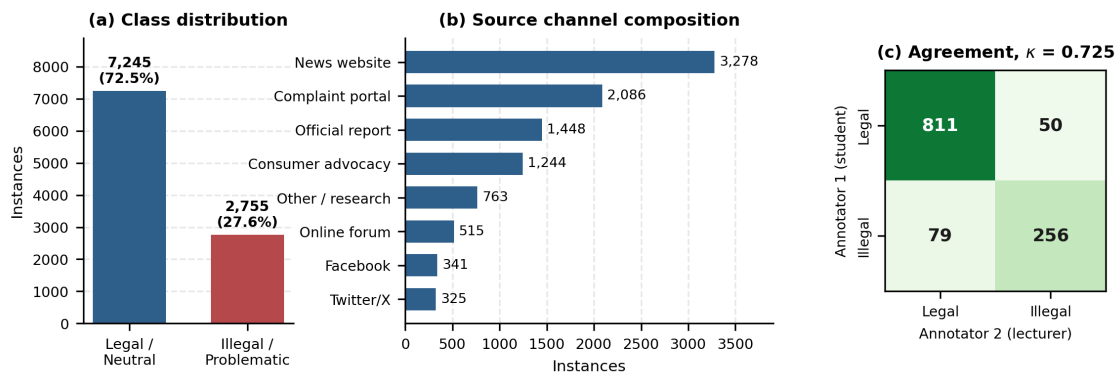


Figure 2. Class distribution (a), source channel composition (b), and inter-annotator agreement on 1,196 doubly labelled texts (c)

The annotators agreed on 1,067 of the 1,196 texts, corresponding to 89.21% observed agreement and a Cohen's kappa of 0.725. The per-partition kappa values were 0.716 and 0.735, respectively, indicating substantial agreement [22]. Among the 129 disagreements, 76 cases were assigned to class 1 after reconciliation, while the remaining 53 cases were

assigned to class 0. The resulting labels were used as the ground truth for the annotated subset. Figure 2 presents the class distribution, source-channel composition, and inter-annotator agreement for the doubly labelled subset.

2.3 Preprocessing and partitioning

Text preprocessing followed six ordered steps: (1) case folding, (2) removal of URLs, mentions, and hashtags, (3) removal of digits and punctuation, (4) domain-specific slang substitution, (5) stopword removal, and (6) stemming. The fixed slang dictionary contained 27 lending-specific terms, including *galbay*, which was normalised to *gagal bayar*, *DC* to *debt collector*, and *norek* to *nomor rekening*. Stopword removal and stemming were performed using Sastrawi, which implements the Nazief-Adriani stemming algorithm. Stemming reduces lexical sparsity but may also remove affixes that carry semantic information [23], [24]. The potential effect of this operation on pretrained subword representations is therefore considered in the discussion of model performance.

Duplicate records were removed before partitioning. Because the corpus was collected from heterogeneous public sources, repeated or near-duplicate content describing the same news event, complaint, or report may still occur. Such potential source-level overlap was considered as a limitation of the partitioning procedure, particularly because exact duplicate removal cannot fully prevent semantic duplication across related reports.

After preprocessing, stratified sampling divided the corpus into 70% training, 15% validation, and 15% test data, corresponding to 7,000, 1,500, and 1,500 instances, respectively. The class ratio was preserved at 27.56% positive in the training partition and 27.53% in both validation and test partitions. The test set consisted of 1,087 negative and 413 positive instances. The random seed was fixed at 42 to ensure reproducibility. Model selection was based exclusively on validation macro F1.

2.4 Models, training configuration, and metrics

A TF-IDF representation limited to 5,000 features with balanced logistic regression was used as the non-neural reference model. This baseline approximates lexical screening because its representation is based on term-level statistics rather than contextual relationships. Three pretrained transformer checkpoints were subsequently fine-tuned: *bert-base-multilingual-cased*, *indobenchmark/indobert-base-p2*, and

indolem/indobertweet-base-uncased. Each model used a randomly initialised two-class classification head operating on the pooled $[CLS]$ representation. Because the illegal class represented approximately 27.6% of the corpus, weighted cross-entropy was used to reduce the tendency of the models to favour the majority class. The class weight for class cc was calculated as Equation 3.

$$w_c = \frac{N}{2n_c} \quad (3)$$

where N is the total number of training instances and n_c is the number of instances belonging to class c . This calculation produced class weights of 0.690 for the legal class and 1.814 for the illegal class. The weighted cross-entropy loss for a mini-batch was defined as Equation 4.

$$\mathcal{L}_{wce} = -\frac{1}{B} \sum_{i=1}^B w_{y_i} \log(p_{i,y_i}) \quad (4)$$

where B is the batch size, y_i is the true class of instance i , w_{y_i} is the corresponding class weight, and p_{i,y_i} is the predicted probability assigned to the true class. Weighting the loss in this manner gives greater contribution to errors involving the minority illegal class. The fine-tuning configuration used for all three transformer models is summarised in Table 3. The same hyperparameters were applied to all models to isolate the effect of the pretrained representation. The learning rate was 2×10^{-5} , batch size was 16, and training was conducted for three epochs using AdamW with a weight decay of 0.01. The maximum sequence length was 128 tokens, mixed-precision *FP16* was enabled, and the random seed was fixed at 42. The model with the highest validation macro F1 was selected for final evaluation.

Table 3. Fine-tuning configuration applied identically to all three transformer models

Parameter	Value	Parameter	Value
Learning rate	2×10^{-5}	Maximum sequence length	128 tokens
Batch size	16	Optimiser	AdamW

Parameter	Value	Parameter	Value
Training epochs	3	Weight decay	0.01
Loss function	Weighted cross-entropy	Mixed precision	FP16
Class weights	0.690 / 1.814	Model selection	Best validation macro F1
Evaluation strategy	Per epoch	Random seed	42

Accuracy was not used as the sole evaluation criterion because a classifier that always predicts the majority class would already obtain approximately 72.4% accuracy. Performance was therefore evaluated using macro-averaged precision, recall, and F1-score, giving equal importance to both classes. For class c , precision, recall, and F1-score were calculated from true positives (TP), false positives (FP), and false negatives (FN) as Equation 5-7.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (5)$$

$$F1_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (6)$$

$$\text{Macro-F1} = \frac{1}{|L|} \sum_{c \in L} F1_c \quad (7)$$

In addition to the classification metrics, ROC-AUC was used to assess the model's threshold-independent discrimination ability. ROC-AUC represents the area under the receiver operating characteristic curve, which relates the true positive rate (TPR) to the false positive rate (FPR) across different classification thresholds. It can be calculated using the trapezoidal approximation as Equation 8.

$$AUC = \sum_{i=1}^{m-1} (FPR_{i+1} - FPR_i) \frac{TPR_{i+1} + TPR_i}{2} \quad (8)$$

where TPR_i and FPR_i denote the true positive rate and false positive rate at the i -th classification threshold, respectively, and m is the number of points on the ROC curve.

A higher ROC-AUC indicates better discrimination between legal and illegal content across classification thresholds. In this study, ROC-AUC and average precision were reported only for models for which prediction probabilities were retained.

3. RESULTS AND DISCUSSION

3.1 Performance across representations

Table 4 summarises the test-set performance of the TF-IDF baseline and the three fine-tuned transformer models on the same 1,500 held-out instances. Figure 3 provides a visual comparison of macro-averaged performance and performance for the illegal class. All three transformer models outperform the lexical baseline across the reported classification metrics. Multilingual BERT achieves the highest macro F1-score of 0.9616, followed by IndoBERT at 0.9582 and IndoBERTweet at 0.9449, compared with 0.8974 for TF-IDF with logistic regression. The corresponding improvements over the baseline are 6.42, 6.08, and 4.75 percentage points, respectively.

The improvement is particularly important for the minority illegal class. As shown in Table 4 and Figure 3(b), the illegal-class F1-score increases from 0.8545 for the lexical baseline to 0.9450 for multilingual BERT, representing a 9.05-point improvement. IndoBERT and IndoBERTweet also improve the illegal-class F1-score to 0.9395 and 0.9209, respectively. These results indicate that contextual representations provide a substantial advantage over lexical features for identifying content associated with the target class.

Table 4. Consolidated test-set performance (n = 1,500; illegal class n = 413)

Model	Acc.	Prec.	Rec.	F1	F1 (ill.)	FP	FN	ROC-AUC	AP
TF-IDF + LogReg	0.9153	0.8862	0.9116	0.8974	0.8545	87	40	0.9792	0.9517
BERT Multilingual	0.9687	0.9528	0.9716	0.9616	0.9450	38	9	–	–
IndoBERT	0.9667	0.9582	0.9582	0.9582	0.9395	25	25	0.9922	0.9819
IndoBERTweet	0.9553	0.9385	0.9519	0.9449	0.9209	44	23	–	–

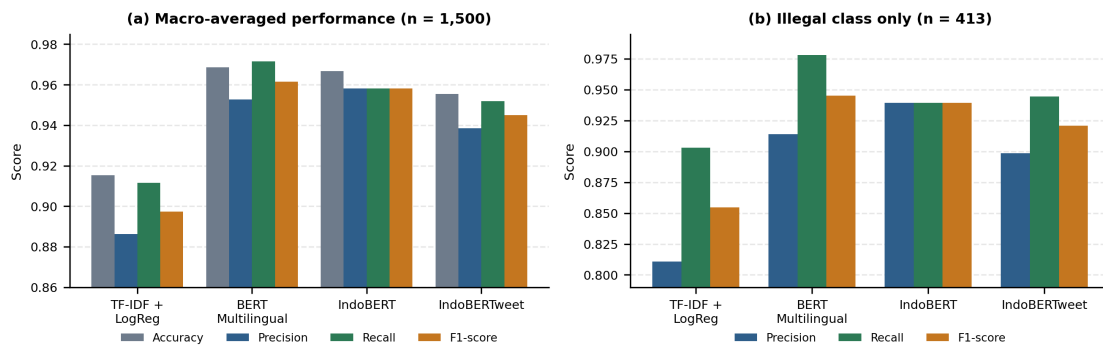


Figure 3. Macro-averaged performance (a) and performance restricted to the illegal class (b)

Although multilingual BERT obtains the highest macro F1-score, its advantage over IndoBERT is small, with a difference of only 0.0034. Because each model was fine-tuned using a single fixed random seed, this difference should not be interpreted as evidence of statistically significant superiority. Instead, the results indicate near-parity between multilingual BERT and IndoBERT on the present corpus. This finding suggests that both multilingual and Indonesian-specific contextual representations can effectively capture the linguistic patterns required for the classification task.

3.2 Error structure

The three transformer models exhibit different error profiles when false positives and false negatives are considered separately. Figure 4 presents the confusion matrices, error composition, and macro-F1 improvement over the lexical baseline. Multilingual BERT misses only 9 of the 413 illegal-class instances, corresponding to a false-negative rate of 2.18%, but produces 38 false positives. IndoBERT produces the fewest false positives, with 25, while IndoBERTweet produces 44 false positives and 23 false negatives. In comparison, the TF-IDF baseline misses 40 illegal instances and incorrectly classifies 87 legal instances as illegal.

Compared with the TF-IDF baseline, multilingual BERT reduces missed illegal instances from 40 to 9, corresponding to a 77.5% reduction, while false positives decrease from 87 to 38, corresponding to a 56.3% reduction. This result indicates that the contextual model provides a substantial improvement in both types of classification error on the held-out test set. The reduction in false negatives is particularly relevant because missed harmful content may reduce the effectiveness of regulatory screening.

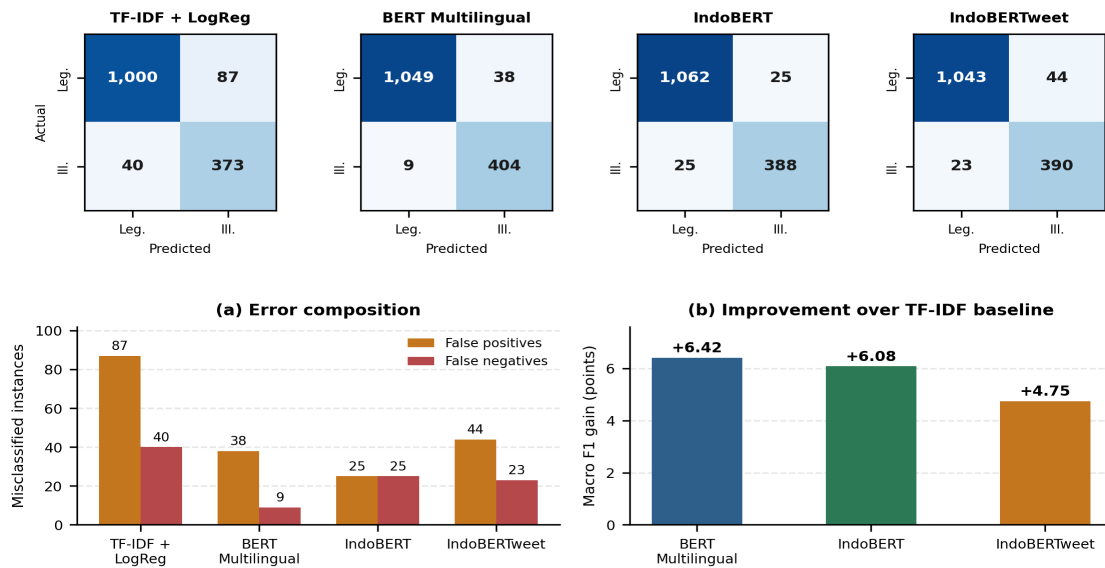


Figure 4. Confusion matrices for all four models (top), error composition (a), and macro F1 gain over the baseline (b)

The error profiles also reveal a precision-recall trade-off between multilingual BERT and IndoBERT. Multilingual BERT achieves higher recall for the illegal class, whereas IndoBERT produces fewer false positives. Thus, multilingual BERT appears preferable when broader coverage is prioritised, while IndoBERT is preferable when reducing false alarms is more important. These differences are relevant to the design of regulatory information systems, where false negatives and false positives may have different operational consequences. However, these observations should be interpreted as model-specific performance characteristics on the present test set rather than evidence of superiority in an operational deployment environment. Further evaluation using repeated random seeds, threshold analysis for all transformer models, and source-stratified or deployment-oriented test sets would be required to establish the stability of these error profiles.

3.3 Threshold-independent discrimination

The results above describe model performance at the default classification threshold of 0.50. Figure 5 further examines model behaviour beyond this single operating point using ROC and precision-recall curves, a threshold sweep, and the fine-tuning trajectory. Because prediction probabilities were retained only for the TF-IDF baseline and IndoBERT, threshold-independent analysis could not be performed consistently for all four models.

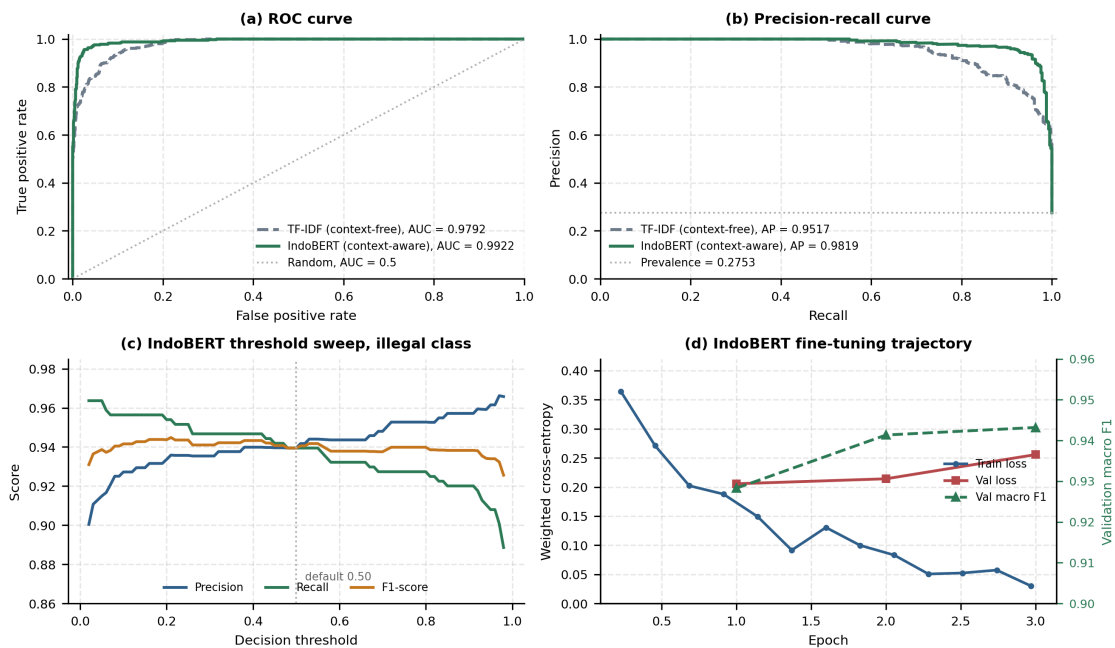


Figure 5. ROC (a) and precision-recall (b) curves, IndoBERT threshold sweep on the illegal class (c), and fine-tuning trajectory (d)

For the models with retained prediction probabilities, IndoBERT achieves a ROC-AUC of 0.9922, compared with 0.9792 for the TF-IDF baseline. Its average precision is also higher, at 0.9819 compared with 0.9517. The precision-recall comparison is particularly relevant because the illegal class represents 27.53% of the test set. Both models operate well above this prevalence level, while IndoBERT maintains stronger precision across the higher-recall region. The IndoBERT threshold sweep in Figure 5(c) shows that macro F1 remains relatively stable across a broad central range of thresholds. The highest observed macro F1 is 0.9448 near a threshold of 0.21, compared with 0.9395 at the default threshold of 0.50. The relatively flat region is more informative than the small difference between the peak and default operating points, indicating that the decision threshold can be adjusted according to the desired balance between recall and review workload.

The threshold analysis should nevertheless be interpreted specifically for IndoBERT and should not be used to infer threshold behaviour for multilingual BERT. In particular, although multilingual BERT achieves the highest recall on the test set, its prediction probabilities were not retained in the current experiment. Therefore, ROC-AUC, average precision, and threshold-sensitivity comparisons cannot be reported for multilingual BERT without rerunning the model inference and retaining its probability outputs.

Figure 5(d) shows an apparent divergence between optimisation loss and classification performance. Training loss decreases from 0.364 to 0.030, while validation loss increases from 0.2057 to 0.2558. At the same time, validation macro F1 improves from 0.9283 to 0.9432. This pattern does not necessarily indicate a deterioration in classification quality. Weighted cross-entropy evaluates the quality of predicted probabilities and penalises confident errors strongly. Consequently, a model can become more confident in a subset of incorrectly classified validation instances and thereby increase its average cross-entropy loss, while simultaneously correcting more classification decisions overall and improving macro F1. This distinction explains why validation macro F1 was selected as the model-selection criterion rather than validation loss, as specified in Table 3. Selecting the model according to loss alone would have favoured an earlier training state, whereas the macro-F1 criterion retained the later model with higher classification performance.

3.4 Fine-Tuning trajectory

Table 5 presents selected test instances in which the TF-IDF baseline and IndoBERT produce different predictions. The examples illustrate why contextual representations can behave differently from lexical representations, particularly when lending-related vocabulary appears in both legal and illegal contexts. The first example contains lending-related terms but explicitly states that the speaker did not provide an OTP. The lexical baseline predicts the instance as illegal with a probability of 0.770, whereas IndoBERT predicts the correct legal class with a probability of 0.995. This indicates that the contextual model can incorporate negation and the relationship between the lending-related terms and the actual assertion in the sentence.

Table 5. Selected test instances where the two representations disagree

Text (abbreviated)	True	TF-IDF	IndoBERT
... kode OTP dari pinjol lain, yang untungnya tidak saya berikan kepada siapa pun.	Legal	Illegal (0.770)	Legal (0.005)
... utang Rp 500.000 dari aplikasi pinjaman online di platform ilegal akan berubah menjadi mimpi buruk.	Legal	Illegal (0.617)	Legal (0.000)

Text (abbreviated)	True	TF-IDF	IndoBERT
<i>Kantor OJK Malang menerima 117 aduan terkait pinjaman online ilegal.</i>	Legal	Illegal (0.760)	Legal (0.003)
<i>Saya sangat menyayangkan atas penagihan oleh pihak debt collector.</i>	Illegal	Illegal (0.944)	Illegal (0.981)
<i>... bunga harian 3,1% sangat memberatkan.</i>	Illegal	Illegal (0.876)	Illegal (1.000)
<i>OJK telah menghentikan atau memblokir 1.591 pinjol ilegal.</i>	Legal	Legal (0.476)	Illegal (0.998)

The second and third examples similarly contain expressions such as *pinjaman online* or references to illegal lending but do not themselves provide evidence of illegal activity. The baseline therefore over-relies on the presence of lending vocabulary, whereas IndoBERT is able to distinguish mention of a concept from evidence of the target behaviour. The final example illustrates the opposite error. The statement that OJK had stopped or blocked 1,591 illegal lending platforms is legally classified as a neutral report, but IndoBERT predicts the illegal class with high confidence. The sentence contains highly diagnostic terms such as *pinjol ilegal*, but these occur in an enumerative regulatory context rather than as evidence of an ongoing illegal promotion or harmful activity. This example helps explain why IndoBERT, despite its contextual advantage, still produces 25 false positives.

These examples demonstrate that contextual modelling improves semantic interpretation but does not eliminate ambiguity. In particular, regulatory reports and statistical statements may contain the same highly discriminative vocabulary as illegal promotional or harmful content. Consequently, the remaining false positives are not simply random errors; they arise partly from legitimate contextual overlap between regulatory descriptions and the target class.

3.5 Discussion

The first research objective was to determine whether contextual transformer representations provide an advantage over a lexical representation. The results provide clear evidence in favour of contextual modelling. TF-IDF treats terms largely as independent features, whereas illegal lending content often depends on how common lending terms are combined with expressions of pressure, repetition, unauthorised access, or borrower harm. For example, terms such as *pinjaman*, *bayar*, *aplikasi*, and *kontak* may occur in both classes, making their individual presence insufficient for reliable classification. Self-attention allows transformer models to represent relationships among these terms and their surrounding context [13]. The examples in Table 5 provide direct qualitative evidence of this mechanism, particularly for negation and contextual interpretation. This finding is also consistent with previous evidence that sentence-level contextual semantics can improve classification over sparse representations [4] and with transformer-based improvements reported for phishing [25] and cyberbullying detection [26].

The second research objective concerned whether Indonesian-specific monolingual pretraining provides a measurable advantage over multilingual BERT. The results do not support a clear advantage in this experimental setting. Previous studies have reported advantages for IndoBERT representations [17], and Nugraha et al. [18] reported 98% accuracy for IndoBERT compared with 96% for multilingual BERT on gambling-related comments. In the present study, however, multilingual BERT obtains a macro F1 of 0.9616 compared with 0.9582 for IndoBERT. The difference is only 0.0034, approximately five instances on the 1,500-item test set. Because each model was trained once with a fixed random seed, this small difference should not be interpreted as statistically established superiority. Rather, the result indicates near parity and raises the more important question of why the expected monolingual advantage does not emerge clearly in this corpus.

One explanation is the interaction between preprocessing and pretrained subword representations. Sastrawi stemming removes Indonesian affixes that can carry important morphological information. Forms such as *menagih*, *ditagih*, and *penagihan* are reduced to *tagih*. IndoBERT was pretrained using unstemmed Indonesian text, so the stemmed

input represents a distribution that differs from the text distribution encountered during pretraining. Part of the morphological information encoded by the pretrained vocabulary is therefore discarded before the model processes the input. Multilingual BERT, which segments Indonesian words into smaller WordPiece fragments, may be proportionally less dependent on the original Indonesian morphological forms.

This interpretation is supported by Khairani et al. [20], who found that stopword removal and stemming reduced accuracy for both IndoBERT and IndoBERTweet. It is also consistent with evidence that the effects of stemming on Indonesian classification depend on the linguistic register and representation used [24]. Therefore, preprocessing designed to improve sparse lexical representations should not automatically be assumed to improve pretrained transformers. In this experiment, stemming may have reduced part of the representational advantage expected from Indonesian-specific pretraining.

A second explanation concerns corpus register. News reporting, regulatory documents, and moderated complaint filings account for approximately 68.1% of the corpus, whereas social-media sources account for only 6.7%. This composition reduces the potential advantage of IndoBERTweet, whose pretraining is more closely associated with Twitter language [16]. IndoBERTweet obtains the lowest macro F1-score among the three transformers, at 0.9449. The ordering therefore suggests that model performance is influenced not only by the amount of Indonesian-language pretraining but also by the match between the model's pretraining domain and the register of the target corpus. Previous monolingual advantages reported in informal registers [18] and [27-28] may consequently narrow when the target data consist predominantly of formal Indonesian. This interpretation is consistent with findings that multilingual and monolingual models can show relatively small differences when the target language and domain are adequately represented in multilingual pretraining [19].

The third research objective concerned the operational precision-recall trade-off. Multilingual BERT misses only 9 illegal texts, corresponding to a 2.18% false-negative rate, but produces 38 false positives. In contrast, IndoBERT produces fewer false positives (25) but misses 25 illegal texts. Thus, multilingual BERT is more suitable for broad regulatory triage where recall is prioritised, whereas IndoBERT is preferable when reducing false alarms is more important.

These complementary error profiles motivate a proposed two-stage regulatory deployment architecture in which multilingual BERT performs broad initial screening and IndoBERT provides higher-precision assessment of the candidate set. The outputs should support, rather than replace, human and regulatory review, particularly because legitimate regulatory or news content may contain overlapping vocabulary.

However, this architecture was not evaluated as an actual cascade in the present study. The reported results represent independent classifiers and do not establish the end-to-end precision, recall, or false-negative rate of a sequential system. Therefore, the architecture should be regarded as a proposed deployment concept rather than an empirically validated cascade. Future work should evaluate the complete pipeline, including candidate volume, second-stage errors, and end-to-end performance. The threshold analysis in Figure 5(c) further indicates that the operating point can be adjusted according to the desired balance between coverage and review workload.

3.6 Threats to Validity and Methodological Limitations

Several limitations should be considered when interpreting these findings. First, each transformer model was fine-tuned once using a fixed random seed of 42. Therefore, the 0.0034 macro-F1 difference between multilingual BERT and IndoBERT should not be interpreted as statistically significant. Repeated runs with multiple seeds and significance testing are needed to assess ranking stability.

Second, only 1,196 validation and test instances were independently annotated by two annotators, achieving a Cohen's kappa of 0.725, while the 8,804 training instances were labelled by a single annotator. Thus, annotation noise may remain in the training data, and the evaluation set was not fully independently verified. Future work should extend independent annotation and apply explicit adjudication to establish more robust ground truth.

Third, although exact duplicates were removed, semantically related reports, complaints, or news articles may remain across partitions. Differences in source composition may also introduce source-specific patterns that models can exploit. Future studies should use source-aware or document-grouped partitioning and evaluate performance on unseen sources or domains.

Fourth, the binary scheme combines different harmful or predatory lending practices into one positive class, including coercive collection, unauthorised data access, debt entrapment, financial harm, fraud, and psychological impact. Multi-label annotation could provide more specific risk information. In addition, the current experiments did not evaluate adversarial or obfuscated text, so robustness against deliberate evasion remains untested.

Finally, stopword removal and stemming were applied to all models, although stemming may alter information used by pretrained subword representations. Since no no-stemming ablation was conducted, its effect on transformer performance remains unclear. Moreover, threshold-independent analysis was available only for TF-IDF and IndoBERT because prediction probabilities were not retained for the other models. Future work should compare preprocessing variants, retain probabilities for all models, and evaluate the proposed two-stage architecture as an actual cascade using end-to-end precision, recall, candidate volume, and false-negative analysis.

4. CONCLUSION

This study demonstrates that contextual transformer models provide a substantial advantage over conventional lexical classification for detecting harmful or predatory illegal fintech lending content on Indonesian digital platforms. Using a corpus of 10,000 texts and a 1,500-instance test set, multilingual BERT achieved the highest macro F1-score of 0.9616, followed by IndoBERT at 0.9582 and IndoBERTtweet at 0.9449, compared with 0.8974 for the TF-IDF and logistic regression baseline. Multilingual BERT also missed only 9 of 413 illegal-class instances, representing a 77.5% reduction in missed cases compared with the baseline, while IndoBERT produced the fewest false positives, with 25 cases. The near parity between multilingual BERT and IndoBERT indicates that Indonesian-specific monolingual pretraining did not provide a clear advantage in the present corpus. The results also highlight the importance of preprocessing and corpus characteristics in determining model performance. The different error profiles motivate a proposed two-stage screening architecture in which multilingual BERT performs high-recall initial screening and IndoBERT provides higher-precision assessment. However, this cascade was not directly evaluated in the present study and should therefore be considered a proposed deployment concept rather than a validated architecture, with

final regulatory decisions remaining subject to human review. Future research should strengthen annotation verification, evaluate source-aware data partitioning and multiple random seeds, compare preprocessing with and without stemming, assess robustness against adversarial and obfuscated text, and retain prediction probabilities for all models. Further evaluation of the proposed two-stage cascade using end-to-end precision, recall, and false-negative analysis is also needed.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Science, and Technology through the Research and Community Service Program, Fiscal Year 2026 (Grant Numbers This work was supported by the Ministry of Higher Education, Science, and Technology through the Research and Community Service Program, Fiscal Year 2026 (Grant Numbers 283/C3/DT.05.00/PL-BARU/2026 and 1553/LL4/PG/2026, 373a/14.LPPM/PPM/IV/2026). The authors also gratefully acknowledge the support provided by Prasetya Mandiri Polytechnic.

REFERENCES

- [1] E. F. Santika, "Thousands of illegal online lending and investment institutions shut down by OJK by June 2025," *Databoks*, Katadata Media Network, Jul. 8, 2025.
- [2] R. R. Suryono, I. Budi, and B. Purwandari, "Detection of fintech P2P lending issues in Indonesia," *Heliyon*, vol. 7, no. 4, pp. e06782, 2021, doi: 10.1016/j.heliyon.2021.e06782.
- [3] D. K. Wardani, S. P. Nugroho, and A. A. Prabowo, "Pengaruh persepsi etis dan religiusitas terhadap niat UMKM melakukan pinjaman online pada masa Covid-19," *Kajian Bisnis*, vol. 29, no. 2, pp. 81–92, 2021, doi: 10.32477/jkb.v29i2.296.
- [4] M. Pan, J. Wang, J. X. Huang, A. J. Huang, Q. Chen, and J. Chen, "A probabilistic framework for integrating sentence-level semantics via BERT into pseudo-relevance feedback," *Inf. Process. Manag.*, vol. 59, no. 1, pp. 102734, 2022, doi: 10.1016/j.ipm.2021.102734.
- [5] R. B. Perdana, Ardin, I. Budi, A. B. Santoso, A. Ramadiah, and P. K. Putra, "Detecting online gambling promotions on Indonesian Twitter using text mining algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 8, pp. 942–949, 2024, doi: 10.14569/IJACSA.2024.0150893.

- [6] B. Stojanovic, J. Bozic, K. Hofer-Schmitz, K. Nahrgang, A. Weber, and A. Badii, "Follow the trail: Machine learning for fraud detection in fintech applications," *Sensors*, vol. 21, no. 5, pp. 1594, 2021, doi: 10.3390/s21051594.
- [7] A. Muzakir, K. Adi, and R. Kusumaningrum, "Short text classification based on hybrid semantic expansion and bidirectional GRU (BiGRU) based method to improve hate speech detection," *Rev. d'Intelligence Artif.*, vol. 37, no. 6, pp. 1471–1481, 2023, doi: 10.18280/ria.370611.
- [8] A. Muzakir, U. Suriani, and U. Ependi, "A hybrid semantic enrichment approach for multi-label toxic speech detection," *Int. J. Saf. Secur. Eng.*, vol. 15, no. 3, pp. 509–520, 2025, doi: 10.18280/ijssse.150310.
- [9] A. Muzakir and U. Suriani, "Model deteksi berita palsu menggunakan pendekatan bidirectional long short term memory (BiLSTM)," *J. Comput. Inf. Syst. Ampera*, vol. 4, no. 2, pp. 93–105, 2023.
- [10] U. Suriani, "Penerapan data mining untuk memprediksi tingkat kelulusan mahasiswa menggunakan algoritma decision tree C4.5," *J. Comput. Inf. Syst. Ampera*, vol. 4, no. 2, pp. 55–65, 2023.
- [11] U. Suriani, I. P. Ninditama, and W. Syaputra, "Pemodelan prediktif keterlambatan bicara pada balita terkait dengan penggunaan smartphone menggunakan data mining," *J. Inf. Technol. Ampera*, vol. 5, no. 1, pp. 129–138, 2024.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter Association For Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)*, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.
- [14] B. Willie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 1st Conf. Asia-Pacific Chapter Association For Computational Linguistics and 10th Int. Joint Conf. Natural Language Processing*, Suzhou, China, 2020, pp. 843–857, doi: 10.18653/v1/2020.aacl-main.85.

- [15] 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 Int. Conf. Computational Linguistics*, Barcelona, Spain, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66..
- [16] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, 2021, pp. 10660–10668, doi: 10.18653/v1/2021.emnlp-main.833.
- [17] G. Z. Nabillah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.
- [18] S. A. Nugraha, C. Lestari, K. B. Sanjaya, R. A. Naya, and J. Jolie, "Comparative analysis of IndoBERT and mBERT for online gambling comment detection in Indonesian social media," *J. Tek. Inform.*, vol. 7, no. 2, pp. 1931–1943, 2026, doi: 10.52436/1.jutif.2026.7.2.5677.
- [19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
- [20] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh tahapan preprocessing terhadap model IndoBERT dan IndoBERTweet untuk mendeteksi emosi pada komentar akun berita Instagram," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 887–894, 2024, doi: 10.25126/jtiik.1148315.
- [21] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
- [22] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [23] S. S. Sarica and J. Luo, "Stopwords in technical language processing," *PLoS One*, vol. 16, no. 8, 2021, e0254937, doi: 10.1371/journal.pone.0254937.
- [24] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *J. Big Data*, vol. 8, no. 1, pp. 26, 2021, doi: 10.1186/s40537-021-00413-1.

- [25] E. Benavides-Astudillo, W. Fuertes, S. Sanchez-Gordon, D. Nunez-Agurto, and G. Rodriguez-Galan, "A phishing-attack-detection model using natural language processing and deep learning," *Appl. Sci.*, vol. 13, no. 9, pp. 5275, 2023, doi: 10.3390/app13095275.
- [26] K. A. Adriana and E. B. Setiawan, "Enhancing cyberbullying detection with a CNN-GRU hybrid model, Word2Vec, and attention mechanism," *J. Tek. Inform.*, vol. 6, no. 3, pp. 1113–1130, 2025, doi: 10.52436/1.jutif.2025.6.3.4176.
- [27] F. Alvin and N. A. S. Winarsih, "Perbandingan kinerja model IndoBERT, IndoBERTweet, dan algoritma klasik pada analisis sentimen isu Indonesia Gelap," *Build. Informatics, Technol. Sci.*, vol. 7, no. 3, pp. 1601–1613, 2025, doi: 10.47065/bits.v7i3.8636.
- [28] K. Kamdan, M. P. Anugrah, M. J. Almutaali, R. Ramdani, and I. L. Kharisma, "Performance analysis of IndoBERT for detection of online gambling promotion in YouTube comments," *Eng. Proc.*, vol. 107, no. 1, pp. 66, 2025, doi: 10.3390/engproc2025107066.