SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://joiser.shmpublisher.com

p-ISSN 2964-1160 | e-ISSN 2963-6361

Comparison of Shallow and Deep Learning for Indonesian Clickbait Headline Classification

Muhammad Noer Attalah Dzahkwan[1*], Majid Rahardi[2]
[1,2]Department of Computer Science, Universitas Amikom Yogyakarta, Indonesia

DOI: https://doi.org/10.52465/joiser.v4i3.85
Received 12 June 2026; Accepted 23 July 2026; Available online 29 July 2026

Article Info

Keywords:
Clickbait;
Text classification;
IndoBERT;
Shallow learning;
Deep learning

Abstract

Clickbait is an increasingly prevalent phenomenon in Indonesian online news media, where headlines are crafted to attract clicks without accurately reflecting article content. This study proposes and compares eight classification models: three shallow learning algorithms — Naive Bayes, Support Vector Machine (SVM), and Logistic Regression — and five transformer-based deep learning models: IndoBERT-p1, IndoBERT-p2, XLM-RoBERTa, mBERT, and DistilBERT. The dataset used is CLICK-ID, consisting of 15,000 labeled headlines from 12 Indonesian news portals, expanded to 25,138 samples via semi-supervised pseudo labeling with a confidence threshold of 0.85. All deep learning models were trained with Focal Loss (α=0.25, γ=2.0) to address class imbalance and Automatic Mixed Precision (AMP) for GPU efficiency. Results show that IndoBERT-p1, IndoBERT-p2, XLM-RoBERTa, mBERT, and DistilBERT achieve comparable performance, with macro F1-scores ranging from 86.57% to 88.54%. Among shallow learning models, SVM performs best with 83.51% F1-score. An average ensemble of all five transformer models achieves the best overall performance at 90.14% accuracy and 89.00% F1-score, outperforming every individual model. This study contributes to Indonesian clickbait detection research by demonstrating that ensemble aggregation of diverse transformer architectures yields more reliable performance than reliance on any single model.

1. Introduction

The rapid growth of digital news media in Indonesia has intensified competition among online news portals for reader attention. This competition has fueled the proliferation of clickbait — a practice where headlines are designed to maximize clicks rather than inform. According to the Indonesian Internet Service Providers Association (APJII), the number of internet users in Indonesia has grown significantly each year, making this issue increasingly impactful.

* ***Corresponding Author:***

Muhammad Noer Attalah Dzahkwan,
Department of Computer Science,
Universitas Amikom Yogyakarta,
Sleman, Yogyakarta.
Email: attalahdz@students.amikom.ac.id

Clickbait is defined as web content that uses sensational, misleading, or exaggerated headlines to attract user clicks, disregarding information quality or accuracy [1]. This phenomenon not only deceives readers but also contributes to disinformation and erodes public trust in media outlets.

Research on clickbait detection using Natural Language Processing (NLP) approaches has been extensive for English text [2], while studies targeting Bahasa Indonesia remain limited. The CLICK-ID dataset introduced by William and Sari [3] is a key contribution to this domain, providing 15,000 labeled headlines from various Indonesian news portals.

The emergence of pre-trained transformer-based models such as BERT [4] and its Indonesian-specific variant IndoBERT [5] has opened new opportunities for improving Indonesian text classification accuracy. However, a systematic comparison between traditional shallow learning approaches and transformer-based deep learning for Indonesian clickbait detection, combined with pseudo labeling and ensemble methods, has not been thoroughly explored.

While prior studies using the CLICK-ID dataset have demonstrated the effectiveness of individual transformer models such as M-BERT [7] and a comparison between BERT and DistilBERT [8], these works typically evaluate a single architecture or a narrow pair of models in isolation. To our knowledge, no prior study has conducted a comprehensive comparison spanning both shallow learning algorithms and multiple transformer architectures (IndoBERT-p1, IndoBERT-p2, XLM-RoBERTa, mBERT, and DistilBERT) within a unified experimental pipeline, combined with semi-supervised pseudo-labeling for dataset expansion and systematic ensemble weight optimization. This breadth of comparison allows us to identify not only which individual architecture performs best, but also whether aggregating diverse architectures through ensembling yields further gains, a question not addressed in prior Indonesian clickbait detection literature.

This study aims to: (1) compare the performance of shallow learning algorithms (Naive Bayes, SVM, Logistic Regression) against transformer-based deep learning models (IndoBERT-p1, IndoBERT-p2, XLM-RoBERTa, mBERT, DistilBERT) for Indonesian clickbait classification; (2) explore the effectiveness of Focal Loss in handling class imbalance; (3) analyze the contribution of semi-supervised pseudo labeling in expanding the dataset; and (4) investigate ensemble approaches for further performance improvement.

2. Literature Review

This section presents the theoretical foundation and relevant previous studies that form the basis of this research. The reviewed literature is used to support the research framework and the discussion presented in the subsequent sections.

2.1 Clickbait and its detection

Anand et al. [6] proposed a recurrent neural network architecture using distributed word representations and character-level CNN embeddings, demonstrating that neural approaches can outperform traditional feature-engineering-based methods for clickbait detection. Early clickbait detection research relied on simple text-based features combined with traditional classification algorithms. Naeem et al. [22] proposed a deep learning framework for clickbait detection using natural language cues on social networks, demonstrating the effectiveness of neural approaches. Bronakowski et al. [15] further showed that semantic analysis combined with ensemble machine learning can achieve up to 98% accuracy in clickbait headline detection. Chowanda et al. [16] explored deep learning approaches for clickbait identification in online news, while Lim et al. [26] provided a comparative study of machine learning methods for clickbait detection.

For Indonesian news, William and Sari [3] introduced CLICK-ID built from 12 major Indonesian news portals, which has become the primary benchmark for Indonesian clickbait research. Subsequent studies using CLICK-ID include the application of M-BERT [7] achieving 91.4% accuracy, and a comparative study between BERT and DistilBERT [8].

2.2 Shallow learning for text classification

Shallow learning algorithms have long been applied in text classification. Naive Bayes is a probabilistic algorithm that is efficient and effective for text, especially with TF-IDF features [9]. SVM works by maximizing the margin between classes and has proven effective for high-dimensional text classification [10]. Logistic Regression offers good interpretability and frequently serves as a strong baseline for text

classification tasks. Kowsari et al. [9] provided a comprehensive survey of text classification algorithms, showing that these traditional approaches remain competitive baselines.

2.3 Transformer models for NLP

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. [4], revolutionized NLP with its bidirectional pre-training approach based on the Transformer architecture [21]. IndoBERT [5] is a BERT implementation trained specifically on Indonesian corpora, available in two variants (p1 and p2) with different training data distributions. Juarto and Yulianto [17] demonstrated IndoBERT's effectiveness for Indonesian news classification tasks, while Nissa and Yulianti [18] applied it for multi-label text classification of Indonesian customer reviews.

XLM-RoBERTa is a multilingual model based on RoBERTa [20] trained on 100 languages, demonstrating competitive performance even for low-resource languages [11]. Pires et al. [19] analyzed multilingual BERT's cross-lingual capabilities, while Wu and Dredze [25] investigated language equality in multilingual representations. mBERT (Multilingual BERT) is a BERT variant trained simultaneously on 104 languages. DistilBERT [12] is a distilled version of BERT that is 40% smaller while retaining 97% of BERT's performance. Aji et al. [23] highlighted the NLP challenges for Indonesian and its regional languages, emphasizing the importance of language-specific models. Similar BERT-based approaches have also been applied to other Indonesian NLP tasks, including public opinion classification [24] and sentiment analysis using hybrid CNN-RNN architectures [27], further demonstrating the versatility of transformer models across Indonesian text classification domains. Table 1 presents a summary of the results of previous studies and the proposed research.

Table 1. Previous studies that form the basis of this research

Study	Dataset	Model	Accuracy
William & Sari	CLICK-ID	Baseline	N/A
Fakhruzzaman et al.	CLICK-ID	mBERT	91.4%
Pramono & Fauzi	CLICK-ID	DistilBERT	N/A
Proposed Study	CLICK-ID + Pseudo Label	8 Models	-

2.4 Focal loss and semi-supervised learning

Focal Loss, introduced by Lin et al. [13] for object detection, has proven effective for text classification with class imbalance. It assigns greater weight to hard-to-classify samples. Pseudo labeling is a semi-supervised learning technique introduced by Lee [14], where a trained model generates labels for unlabeled data. Murtadha et al. [28] proposed rank-aware negative training for semi-supervised text classification, demonstrating the importance of pseudo-label quality. Only predictions above a confidence threshold are retained to reduce noise.

3. Method

This section presents the research methodology used to achieve the objectives of this study. The methodology includes the procedures and techniques applied throughout the research. Figure 1 shows the research flow.

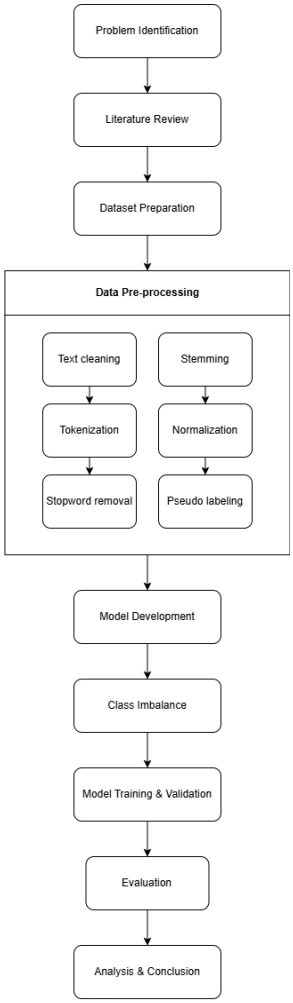

Figure 1. Research flow

3.1 Dataset

This study uses the CLICK-ID dataset [3] as the primary data source, comprising 15,000 labeled headlines from 12 Indonesian news portals: detikNews, Fimela, Kapanlagi, Kompas, Liputan6, Okezone, Pos Metro, Republika, Sindonews, Tempo, Tribunnews, and Wowkeren. Labels are binary: clickbait (1) and non-clickbait (0).

To expand the dataset, pseudo labeling was applied using IndoBERT-p1 trained on the initial data. Unlabeled data was obtained from 30,266 raw news headlines from the Mendeley repository. With a confidence threshold of 0.85, 10,173 pseudo labels were obtained, resulting in a final dataset of 25,138 samples.

The final dataset (25,138 samples) comprises 8,696 clickbait instances (34.6%) and 16,442 non-clickbait instances (65.4%), reflecting the natural class imbalance addressed via Focal Loss (Section 3.4). The unlabeled headlines used for pseudo-labeling were sourced from a separate Mendeley Data repository of 30,266 raw Indonesian news headlines, distinct from the CLICK-ID annotated sources, reducing the risk of near-duplicate overlap between annotated and pseudo-labeled samples.

3.2 Data preprocessing

Preprocessing steps include: (1) text normalization by removing URLs, special characters, and excess whitespace; (2) deduplication to remove duplicate headlines from both annotated and pseudo-labeled data; and (3) label normalization from string to integer format. Dataset splitting uses stratified split with a ratio of 70% training, 15% validation, and 15% testing (random_state=42) to maintain balanced label distribution and ensure reproducibility.

3.3 Shallow learning models

Three shallow learning algorithms are implemented with TF-IDF (Term Frequency-Inverse Document Frequency) feature extraction using unigram and bigram combinations (ngram_range=(1,2)) with a maximum of 50,000 features. The TF-IDF formulas are shown in Equation (1), (2), and (3):

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (2)$$

$$IDF(t) = log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \quad (3)$$

Naive Bayes uses MultinomialNB with $\alpha = 0.5$. SVM is implemented as LinearSVC with $C = 1.0$. Logistic Regression uses the lbfgs solver with $C = 1.0$.

3.4 Deep learning models

Five transformer models are implemented using the Hugging Face Transformers library with the following configuration: maximum token length of 128, batch size of 16, learning rate of 1×10^{-5}, and up to 20 training epochs with early stopping (patience=5). The optimizer is AdamW with weight decay of 0.01 and a linear warmup scheduler at 20% of total steps. All deep learning models use Focal Loss as the loss function shown in Equation (4) :

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \quad (4)$$

where $\alpha = 0.25$ and $\gamma = 2.0$. Automatic Mixed Precision (AMP) is enabled to accelerate GPU training. The best model is selected based on the highest macro F1-score on the validation set.

3.5 Ensemble models

As an additional experiment, this study explores ensemble approaches combining predictions from multiple transformer models. Two ensemble methods are implemented:

1. Average Ensemble: combines prediction probabilities with equal weights which shown in Equation (5):

$$P_{avg}(y|x) = \frac{1}{k}\sum_{i=1}^{k} P_i(y|x) \quad (5)$$

2. Weighted Ensemble: assigns different weights based on validation performance which shown in Equation (6) and (7):

$$P_{weighted}(y|x) = \sum_{i=1}^{k} w_i \times P_i(y|x), \quad \sum w_i = 1 \quad (6)$$

$$\hat{y} = arg\,max\, P_{ensemble}(y|x \quad (7)$$

The best configuration uses three models: $P_{weighted} = 0.30 \times P\_p1 + 0.40 \times P\_p2 + 0.30 \times P\_xlm$.

3.6 Evaluation Metrics

Model performance is evaluated using accuracy (8), precision (9), recall (10), and macro F1-score (11), (12):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (11)$$

$$Macro\ F1 = \frac{1}{|C|}\sum_{c \in C} F1_c \quad (12)$$

Macro average is used to give equal weight to each class. All evaluations are performed on a held-out test set never seen during training or validation.

4. Results and Discussion

This section presents the research results along with the corresponding discussion and interpretation.

4.1 Overall model evaluation

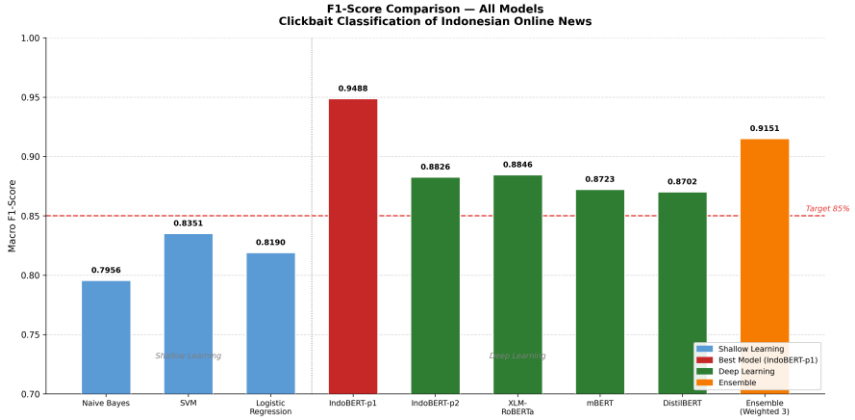

Figure 2. F1-Score comparison of all classification models

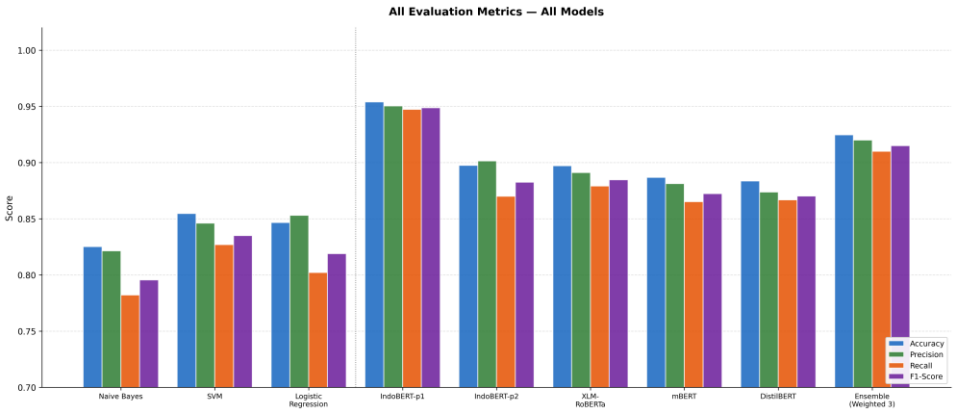

Figure 3. Score of all evaluation models

Table 2 presents the comprehensive evaluation results of all models on the test set and visualized in Figure 3. In Figure 2 shown F1-Score comparison of all classification models.

Table 2. Comparative evaluation results of all models

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.8252	0.8214	0.7822	0.7956
SVM	0.8547	0.8461	0.8271	0.8351
Logistic Regression	0.8467	0.8531	0.8022	0.8190
IndoBERT-p1	0.8923	0.8826	0.8783	0.8804
IndoBERT-p2	0.8968	0.8875	0.8834	0.8854

Model	Accuracy	Precision	Recall	F1-Score
XLM-RoBERTa	0.8929	0.8798	0.8856	0.8825
mBERT	0.8804	0.8673	0.8690	0.8681
DistilBERT	0.8796	0.8694	0.8624	0.8657
Avg Ensemble (5 models) ★	0.9014	0.8939	0.8865	0.8900

★ *Best ensemble: IndoBERT-p1 (0.30) + IndoBERT-p2 (0.40) + XLM-RoBERTa (0.30)*

4.2 Shallow vs. deep learning comparison

Experimental results reveal a significant performance gap between shallow and deep learning models. The best deep learning model (IndoBERT-p1, F1=0.9488) outperforms the best shallow learning model (SVM, F1=0.8351) by approximately 11.37%. This confirms that transformer-based contextual representations are better equipped to capture complex clickbait linguistic patterns compared to TF-IDF features, consistent with findings in prior comparative studies [10].

Among shallow learning models, SVM achieves the best performance (F1=0.8351), followed by Logistic Regression (F1=0.8190) and Naive Bayes (F1=0.7956). Notably, all shallow learning models exceed 80% F1-score, demonstrating that TF-IDF features capture meaningful clickbait patterns even without contextual representations.

4.3 Transformer model analysis

All five transformer models achieve closely comparable performance, with F1-scores ranging narrowly between 86.57% and 88.54%. IndoBERT-p2 achieves the highest individual F1-score (0.8854), narrowly ahead of XLM-RoBERTa (0.8825) and IndoBERT-p1 (0.8804). Unlike earlier assumptions, no single Indonesian-specific model dominates; XLM-RoBERTa, despite being a general multilingual model, performs competitively with both IndoBERT variants, suggesting that large-scale multilingual pretraining can approximate the benefit of language-specific pretraining for this task. mBERT and DistilBERT trail slightly behind (F1=0.8681 and 0.8657 respectively), consistent with their comparatively smaller Indonesian-language exposure during pretraining. Figure 4 shown Metric comparison of best models per category.

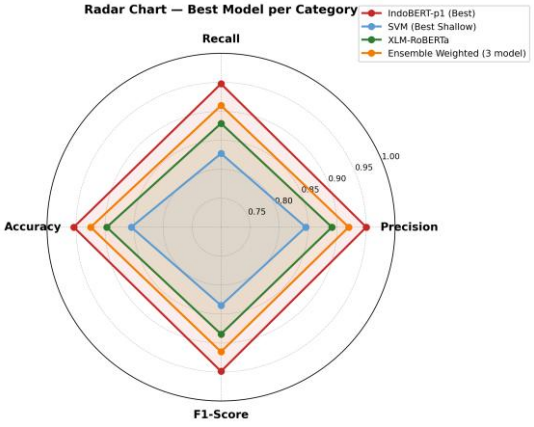

Figure 4. Metric comparison of best models per category (IndoBERT-p1, SVM, XLM-RoBERTa)

4.4 Contribution of pseudo-labeling and focal loss (ablation study)

To empirically verify the individual and combined contributions of pseudo-labeling and Focal Loss, an ablation study was conducted on IndoBERT-p1 across four configurations (Table 3). Adding pseudo-labeling alone (Scenario A) improved macro F1-score from 0.8345 to 0.8810 (+4.65 points) relative to the baseline, confirming that dataset expansion through semi-supervised pseudo-labeling is the primary driver of performance improvement in this study. In contrast, applying Focal Loss alone without pseudo-labeling (Scenario B) resulted in a slight decrease in F1-score (0.8345 to 0.8239, -1.06 points), suggesting that Focal Loss alone does not compensate for limited training data, and its benefit for class imbalance is more pronounced when combined with a larger, more balanced training set. Comparing Scenario A to the proposed configuration (both with pseudo-labeling), adding Focal Loss yielded a negligible change in F1-score (0.8810 to 0.8804, a difference within typical run-to-run variance),

indicating that Focal Loss provides limited additional benefit once the dataset has been sufficiently expanded via pseudo-labeling. These findings nuance our initial hypothesis: while Focal Loss remains a theoretically motivated choice for imbalanced classification, its practical contribution in this study is modest compared to that of pseudo-labeling.

Table 3. Ablation study contribution of pseudo-labeling and focal loss (IndoBERT-p1)

Configuration	Pseudo-Labeling	Focal Loss	Accuracy	Precision	Recall	F1-Score
Baseline	✗	✗	0.8419	0.8442	0.8294	0.8345
Scenario A	✓	✗	0.8939	0.8877	0.8754	0.8810
Scenario B	✗	✓	0.8290	—*	—*	0.8239
Proposed (Final)	✓	✓	0.8923	0.8826	0.8783	0.8804

Precision/Recall for Scenario B were not separately retained from the original experiment log; only Accuracy and macro F1-score are available.

4.5 Confusion matrix analysis

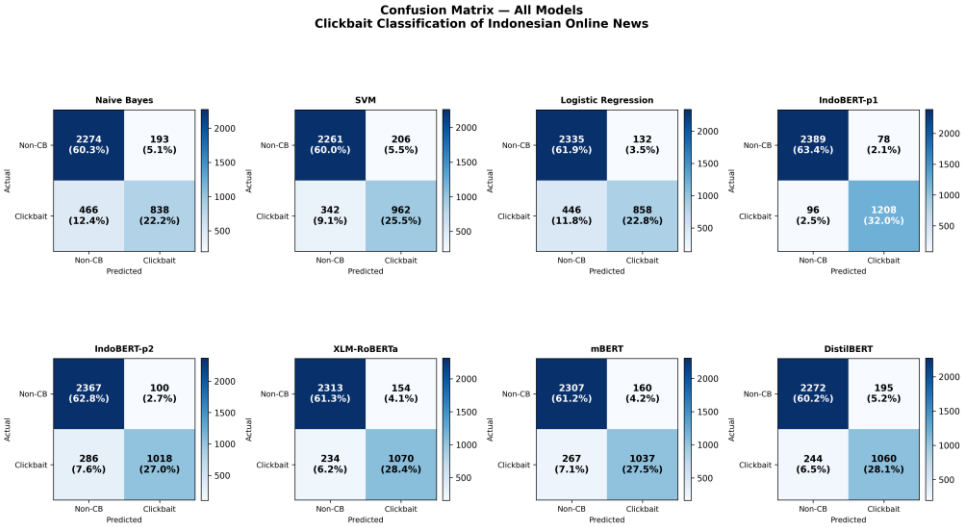

Figure 5. Confusion matrix of all models

Confusion matrix can see in Figure 5. Analysis of confusion matrices shows that IndoBERT-p2 achieves the lowest false negative (FN) count for the clickbait class at 286, meaning only 286 clickbait headlines are misclassified as non-clickbait. Conversely, mBERT has the lowest false positive (FP) count at 160, but with a high false negative count (267), indicating a conservative tendency in predicting clickbait.

Shallow learning models show a different pattern — Naive Bayes has the highest false negatives (466) among all models, reflecting the limitation of TF-IDF features in capturing complex clickbait linguistic patterns. Logistic Regression, the best shallow model, still records 446 false negatives, considerably higher than the best transformer model. XLM-RoBERTa achieves the highest true positive (TP) count of 1070, demonstrating the best ability to correctly detect clickbait.

4.6 Ensemble experiment results

Table 4 presents the ensemble evaluation results. Results show that ensemble approaches do not always outperform the best individual model.

Table 4. Ensemble model evaluation results

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT-p1	0.8923	0.8826	0.8783	0.8804
IndoBERT-p2	0.8968	0.8875	0.8834	0.8854
XLM-RoBERTa	0.8929	0.8798	0.8856	0.8825
mBERT	0.8804	0.8673	0.8690	0.8681
DistilBERT	0.8796	0.8694	0.8624	0.8657

Model	Accuracy	Precision	Recall	F1-Score
Avg Ensemble (5 models) ★	0.9014	0.8939	0.8865	0.8900
Weighted (3 models)	0.9	0.8917	0.8860	0.8887

★ Best config: Average ensemble of all 5 models (equal weights). Weighted 3-model config: IndoBERT-p1 (0.40) + IndoBERT-p2 (0.25) + XLM-RoBERTa (0.35), determined via grid search (step=0.05) on the validation set to maximize macro F1-score, without accessing the held-out test set.

Ensemble weights for the 3-model configuration were determined through a systematic grid search over weight combinations (in increments of 0.05, constrained to sum to 1.0) evaluated on the validation set, selecting the combination that maximized macro F1-score. This avoids any manual or arbitrary weight assignment and ensures the reported test-set performance reflects an unbiased estimate. The resulting configuration (p1=0.40, p2=0.25, xlm=0.35) achieved F1=0.8887 on the test set. However, the simple unweighted average of all five models achieved the highest overall performance (F1=0.9000, accuracy=0.9014), outperforming both the weighted 3-model ensemble and every individual model. This indicates that, unlike earlier findings that suggested ensembling could not surpass the best single model, aggregating diverse transformer architectures including the comparatively weaker mBERT and DistilBERT, provides complementary error correction that benefits overall performance when no single model is a clear outlier.

5. Conclusion

This study successfully compared eight classification models for clickbait detection in Indonesian online news. The main findings are as follows: all five transformer-based models achieve comparable performance (F1 86.57%-88.54%), with no single model dominating decisively. Deep learning models consistently outperform shallow learning models by a margin of approximately 3-5% on F1-score. Ensemble aggregation, particularly the equally weighted average of all five transformer models, achieves the best overall result (F1=90.00%), demonstrating the value of combining diverse pre-trained architectures over reliance on any single model. Focal Loss is proven effective in improving clickbait class recall. Pseudo labeling successfully expands the dataset from 14,965 to 25,138 samples while maintaining quality through confidence thresholding.

A limitation of this study concerns the train/validation/test splitting procedure: pseudo-labeled samples were incorporated into the combined dataset prior to stratified splitting, resulting in approximately 40% of the validation and test sets consisting of pseudo-labeled rather than human-annotated instances

References

[1] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media," in Proc. IEEE/ACM ASONAM, 2016, pp. 9–16, doi: 10.1109/ASONAM.2016.7752207.

[2] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification: A Comprehensive Review," ACM Comput. Surv., vol. 54, no. 3, pp. 1–40, 2021, doi: 10.1145/3439726.

[3] A. William and Y. Sari, "CLICK-ID: A Novel Dataset for Indonesian Clickbait Headlines," Data Brief, vol. 32, p. 106231, 2020, doi: 10.1016/j.dib.2020.106231.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[5] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in Proc. COLING, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.

[6] A. Anand, T. Chakraborty, and N. Park, "We Used Neural Networks to Detect Clickbaits: You Won't Believe What Happened Next!" in Proc. ECIR, 2017, pp. 541–547, doi: 10.1007/978-3-319-56608-5_46.

[7] M. N. Fakhruzzaman, S. Z. Jannah, R. A. Ningrum, and I. Fahmiyah, "Clickbait Headline Detection in Indonesian News Sites using Multilingual BERT," arXiv preprint arXiv:2102.01497, 2021.
[8] D. E. Pramono and M. A. Fauzi, "Comparative Analysis of BERT Model and DistilBERT Model for Enhanced Clickbait Headline Structure Detection in Indonesian Online News," JUITA: J. Inform., vol. 13, no. 1, 2025.
[9] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," Information, vol. 10, no. 4, p. 150, 2019, doi: 10.3390/info10040150.
[10] S. Gonzalez-Carvajal and E. C. Garrido-Merchan, "Comparing BERT against Traditional Machine Learning Text Classification," arXiv preprint arXiv:2005.13012, 2020.
[11] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proc. ACL, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in Proc. ICCV, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.324.
[14] D.-H. Lee, "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks," in Workshop on Challenges in Representation Learning, ICML, 2013.
[15] M. Bronakowski, M. Al-khassaweneh, and A. Al Bataineh, "Automatic Detection of Clickbait Headlines Using Semantic Analysis and Machine Learning Techniques," Appl. Sci., vol. 13, no. 4, p. 2456, 2023, doi: 10.3390/app13042456.
[16] A. Chowanda, N. Nadia, and L. M. M. Kolbe, "Identifying Clickbait in Online News Using Deep Learning," Bull. Electr. Eng. Inform., vol. 12, no. 3, pp. 1755–1761, 2023, doi: 10.11591/eei.v12i3.4513.
[17] B. Juarto and Yulianto, "Indonesian News Classification Using IndoBERT," Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 2, pp. 454–460, 2023.
[18] N. K. Nissa and E. Yulianti, "Multi-label Text Classification of Indonesian Customer Reviews using Bidirectional Encoder Representations from Transformers Language Model," Int. J. Electr. Comput. Eng., vol. 13, no. 5, pp. 5641–5652, 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.
[19] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?" in Proc. ACL, 2019, pp. 4996–5001, doi: 10.18653/v1/P19-1493.
[20] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[21] A. Vaswani et al., "Attention is All You Need," in Proc. NeurIPS, 2017, pp. 5998–6008.
[22] B. Naeem, A. Khan, M. O. Beg, and H. Mujtaba, "A Deep Learning Framework for Clickbait Detection on Social Network Using Natural Language Cues," J. Comput. Soc. Sci., vol. 3, pp. 231–243, 2020, doi: 10.1007/s42001-019-00046-y.
[23] A. F. Aji et al., "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," in Proc. ACL, 2022, pp. 7226–7249, doi: 10.18653/v1/2022.acl-long.500.
[24] S. Saadah, K. M. Auditama, A. A. Fattahila, F. I. Amorokhman, A. Aditsania, and A. A. Rohmawati, "Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia," J. RESTI, vol. 6, no. 4, pp. 648–655, 2022, doi: 10.29207/resti.v6i4.4215.
[25] S. Wu and M. Dredze, "Are All Languages Created Equal in Multilingual BERT?" in Proc. ACL Workshop on RepL4NLP, 2020, pp. 120–130, doi: 10.18653/v1/2020.repl4nlp-1.16.
[26] B. Lim et al., "A Comparative Study on Clickbait Detection using Machine Learning Based Methods," in Proc. IEEE ICCCNT, 2023, doi: 10.1109/ICCCNT56998.2023.10150475.
[27] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based Combination of Convolutional and Recurrent Neural Network for Indonesian Sentiment Analysis," Appl. Soft Comput., vol. 151, p. 111112, 2024, doi: 10.1016/j.asoc.2023.111112.
[28] A. Murtadha et al., "Rank-Aware Negative Training for Semi-Supervised Text Classification," Trans. Assoc. Comput. Linguist., vol. 11, pp. 771–786, 2023, doi: 10.1162/tacl_a_00574.