CLICK-ID MULTI: A Multimodal Dataset for Indonesian Clickbait Detection and Benchmarking

Anonymous ACL submission

Abstract

Clickbait headlines attract user attention by exploiting curiosity gaps, often through sensational or misleading phrasing, while not necessarily conveying false information. Although clickbait contributes to the broader misinformation ecosystem, especially when amplified on social media, it remains underexplored in low-resource and multimodal settings. This paper introduces **CLICK-ID MULTI**, a new multimodal dataset for clickbait detection in Indonesian. It extends the original CLICK-ID dataset (William and Sari, 2020) by pairing 5,809 annotated news articles with associated images, enabling the development of multimodal models. Despite its smaller size compared to the original text-only dataset, CLICK-ID MULTI supports models that outperform the best text-only baseline (F1 = 0.7365), achieving F1 scores up to 0.937 through image-text fusion. These findings highlight the importance of multimodal learning and language-specific pretraining for robust clickbait detection in low-resource languages. The dataset and code are publicly available at: https://anonymous.4open.science/r/emnlp-2025-clickid-multi-8466.

1 Introduction

The rapid spread of fake news, amplified by social media, has become a major societal challenge. While automated fake news detection has advanced, most research focuses on high-resource languages, especially English (Wang, 2017). In contrast, low-resource languages (LRLs) suffer from a lack of annotated datasets, limiting the development of effective detection models (Cieri et al., 2016).

While fake news detection has been widely studied, the role of clickbait in spreading misinformation is still not well understood. Although both phenomena exploit exaggerated or misleading headlines to attract attention and drive traffic, often at the expense of journalistic integrity (Chakraborty et al., 2016; Fakhruroji et al., 2023), they differ fundamentally in purpose. Fake news is primarily concerned with spreading false or fabricated content (veracity), whereas clickbait is driven by the intent to provoke curiosity and generate engagement by creating information gaps, even when the content itself is factually accurate (Scott, 2021). Such practices are further reinforced by media logics that prioritize engagement metrics over editorial standards, especially in digital newsrooms (Fakhruroji et al., 2023). This not only harms the credibility of online news but also reduces public trust in the media. Clickbait can spread false or distorted information, especially in political news, where it may influence public opinion and even election outcomes (Chen et al., 2015; Molyneux and Coddington, 2020). The problem is made worse by social media algorithms that promote engaging content, further increasing the reach of misleading headlines (Chen et al., 2015). Because of its role in misinformation, detecting clickbait is essential for improving the filtering of misleading or manipulative content before it spreads widely.

However, most clickbait detection methods focus solely on text, overlooking the visual aspect, particularly images embedded in articles, which may either reinforce or contradict the headline in ways that serve as valuable clues for distinguishing between clickbait and non-clickbait content (Shrestha et al., 2024; Choi et al., 2022; Yu et al., 2024). This highlights the need to explore multimodal approaches that leverage both modalities. Despite this potential, multimodal clickbait detection remains understudied, especially in low-resource languages.

Most existing fake news datasets are designed for English, including LIAR (Wang, 2017) and FakeNewsNet (Shu et al., 2020), while resources for other languages remain scarce. Some efforts have introduced datasets in Mandarin, such as CHEF (Hu et al., 2022), and German, such as FANG-

COVID (Mattern et al., 2021), but these remain limited. Multimodal datasets that combine text and images are even rarer for low-resource languages (LRLs) (Nakamura et al., 2020). In particular, Indonesian, a language spoken by over 200 million people, still lacks large-scale annotated datasets for both fake news and clickbait detection (Isa et al., 2022; Mahendra et al., 2021).

To address this, we introduce refined and enhanced version of the CLICK-ID dataset (William and Sari, 2020), it integrates both textual and visual modalities, enabling multimodal learning. CLICK-ID MULTI contains 5,809 annotated news articles, each paired with an extracted image, supporting multimodal classification approaches.

This study contributes by (1) introducing a multimodal dataset for clickbait detection in Indonesian, (2) benchmarking text-only, image-only, and multimodal classification models, and (3) analyzing multimodal fusion's role in clickbait detection.

In this context, the remaining parts of the text are organized as follows. Section 2 explains how the original CLICK-ID dataset was created and describes the curation process for CLICK-ID MULTI. Section 3 reviews state-of-the-art multimodal approaches. Section 4 outlines the benchmark evaluation setup, while Section 5 presents the results and discussion. Finally, Section 6 provides the conclusion and future directions.

2 CLICK-ID MULTI: a new multimodal dataset

The original CLICK-ID dataset (William and Sari, 2020) was constructed from 12 Indonesian news publishers using dedicated scrapers. From the collected articles, 15,000 headlines were annotated as clickbait or non-clickbait by undergraduate students. Each headline was labeled by three annotators, with the majority vote used to determine the final label. The dataset achieved a moderate inter-annotator agreement, with a Fleiss' Kappa score of 0.42. The study by (William and Sari, 2020) reports accuracy as the primary evaluation metric but omits Precision, Recall, F1-score, and ROC-AUC, which are particularly important in imbalanced classification settings where accuracy alone can be misleading (Puneetha et al., 2025).

For this study, we retrieved the dataset from Kaggle,[1] which includes a *raw* folder (full articles) and an *annotated* folder (titles and labels). Missing identifiers were reconstructed by matching titles and assigning unique IDs based on their news publisher (e.g., `fimela_0`).

To extend the dataset multimodally, we extracted images from the news URLs using `BeautifulSoup`.[2] We hypothesize that visual context may amplify or mitigate clickbait effects, motivating this enrichment. However, not all articles contained usable images, and we observed substantial variability across publishers in HTML structure and image availability. Due to this inconsistency and changes in site architecture over time (2020–2024), a consistent retrieval rate could not be established. Images were considered usable if they corresponded to the main article content (excluding ads, thumbnails, or logos), were directly parsable, and had valid URLs. Articles without such images were excluded from the multimodal extension.

As shown in Table 1, the final dataset contains 5,809 image-text pairs.

Publisher	Non-Clickbait	Clickbait	Images
Fimela	299	371	670
Kapanlagi	438	327	765
Kompas	1,104	328	1,432
Liputan6	606	850	1,456
Okezone	732	754	1,486
Total	**3,179**	**2,630**	**5,809**

Table 1: Publisher statistics and image counts.

Text, code, and exploratory statistics (e.g., average document length, sentence count) are available at `https://anonymous.4open.science/r/emnlp-2025-clickid-multi-8466`. Due to size constraints, image data will be released upon acceptance.

3 Multimodal fake news detection

The primary objective of this work is to contribute to the development of datasets that support research in fake news and clickbait detection. To achieve this, we investigate whether a multimodal detection approach can be effectively designed using the proposed Click-ID MULTI dataset. Given the strong performance of transformer-based models, such as BERT (Devlin, 2018; Szczepański et al., 2021) and RoBERTa (Liu, 2019; Angizeh and Keyvanpour, 2024), in textual classification, as well as deep

[1] `https://www.kaggle.com/datasets/andikawilliam/clickid`

[2] `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`

learning architectures like EfficientNetB0 (Tan and Le, 2019) and ResNet50 (He et al., 2016) in image processing, this work integrates these models to enhance multimodal detection. To further explore more recent methods, we also evaluate vision-language models such as CLIP and BLIP (Li et al., 2023), which offer powerful pretrained representations for image-text understanding. Additionally, we assess the ability of large language models (LLMs), such as LLaMA-3 (Touvron et al., 2023), to classify clickbait in our text-only dataset using prompt-based inference. As headline-only input yielded poor performance, we condition the model on both the headline and the full article content. The full prompt used for this evaluation is provided in Appendix B Specifically, the textual component is processed using transformer-based encoders, while the visual component is analyzed using convolutional neural networks (CNNs) or pretrained vision encoders. The modality fusion is performed at the feature level, where the text and image embeddings are concatenated and passed to a classification layer. This allows the model to learn joint cross-modal interactions. Our experiments compare unimodal baselines (text-only, image-only) against multimodal variants to assess the benefit of visual cues in improving classification performance in a low-resource setting like Indonesian.

4 CLICK-ID MULTI: benchmark evaluation

Let us now outline the experimental setup and describe the baseline as well as recent vision-language and language model architectures. We then present the empirical results and conduct a comparative analysis of different models.

4.1 Experimental settings

The experimental setup considers three modalities: (1) **Text-only**, using BiLSTM, CNN, BERT, and LLaMA-3 (Touvron et al., 2023). LLaMA-3 is evaluated in a zero-shot, prompt-based setting using both headline and article content, as headline-only input resulted in poor performance. For BERT-based models, we fine-tune three variants: Indonesian-pretrained IndoBERT, multilingual BERT, and English BERT. These are denoted respectively as BERT (id), BERT (m), and BERT (en) in the results tables. (2) **Image-only**, using ResNet50, EfficientNetB0, CLIP, and BLIP (Li et al., 2023). ResNet50 is selected for subsequent fusion experiments due to its comparable performance to EfficientNetB0 and lower resource requirements. (3) **Multimodal**, with two types of models: end-to-end pretrained encoders (CLIP, BLIP) and supervised fusion baselines, in which ResNet50 image embeddings are concatenated with text embeddings (from BiLSTM, CNN, or BERT) and passed through a classifier. All models are trained using 5-fold cross-validation. BiLSTM and CNN are trained for 5 epochs (learning rate 1×10^{-3}), BERT-based models for 3 epochs (2×10^{-5}), and image-based or multimodal models for 10 epochs (1×10^{-5}). Although the original CLICK-ID paper (William and Sari, 2020) did not report training epochs, these settings are informed by prior work (Agrawal, 2016) and confirmed via preliminary convergence analysis.

4.2 Evaluation metrics

The performance of the classification was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC, as commonly reported in similar approaches (e.g (Farhadpour et al., 2024)). In each case, standard definitions of these metrics have been used. The complete set of results has been presented in Appendix: in Table 6, Table 7, and Table 8, Table 9.

5 Results

Table 2 reports model performance across text-only, image-only, multimodal, and zero-shot LLM-based settings, using Accuracy and F1-score. Given class imbalance in CLICK-ID MULTI, F1-score is the more informative metric.

Text-only models performed well overall, with BERT (id) achieving the highest F1-score (0.7365), outperforming its multilingual (0.6983) and English (0.6647) counterparts. CNN and BiLSTM models trailed behind, confirming the advantage of transformer-based and language-specific pretraining for Indonesian.

Image-only models performed poorly. Despite achieving moderate accuracy (64%), CLIP and BLIP defaulted to majority-class predictions, resulting in F1-scores of 0.000. Other vision models like ResNet50 and EfficientNet hovered near chance.

Multimodal models significantly outperformed unimodal ones. The best result was obtained by BLIP + BERT (id) with an F1-score of **0.9372**,

followed by CLIP + BERT (id) at 0.9152. This highlights the strength of pairing modern vision-language encoders with language-specific transformers.

LLM-based zero-shot inference with LLaMA 3 (8B) yielded modest results (F1 = 0.5246), suggesting limited utility without task-specific adaptation.

Overall, combining BLIP/CLIP with BERT (id) offers the strongest performance, while image-only inputs and general-purpose LLMs remain inadequate for this task.

Model	Modality	Accuracy	F1 Score
BLIP + BERT (id)	Text+Images	0.9440	$\mathbf{0.9372 \pm 0.1134}$
CLIP + BERT (id)	Text+Images	0.9390	0.9152
BiLSTM-ResNet	Text+Images	0.6615	0.7303
Resnet + BERT (id)	Text+Images	0.8017	0.7279
Resnet + BERT (m)	Text+Images	0.7921	0.7022
Resnet + BERT (en)	Text+Images	0.7653	0.6727
CNN-ResNet	Text+Images	0.5500	0.5519
BERT (id)	Text	0.8062	$\mathbf{0.7279 \pm 0.0070}$
BERT (m)	Text	0.7901	0.6983
CNN	Text	0.7582	0.6732
BERT (en)	Text	0.7749	0.6647
BiLSTM	Text	0.7543	0.6574
LLaMA 3 (8B)	Text	0.5409	0.5246
ResNet50	Images	0.4999	**0.2669**
EfficientNet	Images	0.4999	0.2669
CLIP	Images	0.6402	0.0000
BLIP	Images	0.6402	0.0000

Table 2: Performance metrics across different modalities, sorted by F1 Score within each modality. For text+image fusion, BLIP + BERT (id) performs best. For text-only, BERT (id) achieves the highest F1-score. Among image-only models, ResNet50 and EfficientNet show the highest (though still low) F1-scores. CLIP and BLIP yield zero F1-scores due to majority-class predictions. Full classification metrics are provided in Appendix D.

6 Concluding remarks

This study introduces **CLICK-ID MULTI**, a novel dataset for multimodal clickbait detection in Indonesian, a low-resource language. By combining textual and visual modalities, it enables comprehensive evaluation of unimodal and multimodal models in detecting misleading headlines.

Our experiments show that text-only models consistently outperform image-only models, with the Indonesian-pretrained BERT (BERT (id)) achieving the highest F1 score among all textual approaches. In contrast, image-only models exhibit moderate accuracy but near-zero F1 scores, driven by prediction collapse toward the majority class—highlighting the impact of class imbalance when visual information is used in isolation.

Multimodal **fusion** yields significant improvements, particularly when BERT (id) is paired with vision-language encoders such as CLIP and BLIP. Fusion models like BLIP + BERT (id) and CLIP + BERT (id) significantly outperform the BERT (id) -only baseline ($p < 0.05$, paired t-test), achieving average F1 scores of 0.937 and 0.915, respectively. These results demonstrate that visual features in Indonesian news images are informative and contribute meaningfully to performance when effectively fused with strong textual representations.

While BiLSTM–ResNet fusion led to a notable gain (+0.0729), fusing Resnet50 + BERT (id) resulted in a slight performance decline (-0.0086), suggesting that transformer-based models may require more sophisticated **cross-modal alignment** strategies to fully benefit from visual signals. This points to a need for architectures that explicitly model interactions between modalities rather than relying on naive fusion alone.

A zero-shot evaluation using LLaMA 3 (8B) achieved an F1 score of 0.5246 without fine-tuning, indicating that instruction-tuned LLMs can offer meaningful baselines for low-resource multimodal tasks.

In summary, the findings highlight that: (1) text remains the most reliable modality, (2) image-only models struggle due to class imbalance and lack of discriminative power, and (3) fusion especially when using BERT (id) with modern visual encoders can surpass text-only baselines. Further gains may be possible through improved cross-modal alignment techniques.

Future work should investigate strategies to mitigate the class imbalance affecting image-only models.

7 Limitations

The primary limitation of this study lies in dataset size and completeness. Many news articles lack accompanying images, which constrains the potential of multimodal learning. The dataset contains 5,809 annotated samples—substantially smaller than the 15,000 text-only articles in the original CLICK-ID dataset (William and Sari, 2020), which limits generalizability.

In addition, the quality and relevance of available images may not provide sufficiently discriminative visual cues, weakening the effectiveness of multimodal **fusion**. This is reflected in the per-

formance of image-only models (e.g., CLIP and BLIP), which achieved moderate accuracy but consistently failed to produce non-zero F1 scores due to class imbalance and prediction collapse. Traditional vision models such as ResNet50 and EfficientNet primarily extract low-level features, which may not capture the semantic context necessary for informative fusion with textual inputs.

Another limitation concerns the **cross-modal alignment** between textual and visual representations. For instance, the fusion of BERT (id) with ResNet50 resulted in a slight performance drop (-0.0086), suggesting that simple fusion methods such as feature concatenation may fail to produce aligned representations. In contrast, stronger performance was observed when BERT (id) was paired with pretrained image-language models such as CLIP or BLIP, which offer better cross-modal alignment due to joint pretraining on vision-language tasks.

Finally, as Indonesian is a low-resource language, existing multimodal pretrained models may not be optimized for Indonesian text-image interactions. Future work should consider expanding the multimodal dataset, applying domain-adaptive pretraining, and developing stronger **fusion strategies**. In particular, alignment mechanisms based on cross-modal attention or shared embedding spaces may help reduce modality mismatch and enhance integration quality.

8 Ethical considerations

This study focuses on multimodal clickbait detection using a dataset collected from Indonesian news sources. We acknowledge the importance of ethical considerations in dataset creation, ensuring that all data used complies with fair use policies and is intended solely for research purposes. The dataset does not contain personally identifiable information, and no modifications were made to the original news content that could misrepresent the intent of the sources.

We would like to thank Indonesian news agencies such as Fimela, Kapanlagi, Kompas, Liputan6, and Okezone for providing publicly available news content, which serves as a valuable resource for advancing fake news and clickbait detection research in low-resource languages.

Acknowledgments

After acceptance for publication, all sources of funding and contributions from collaborators will be properly acknowledged.

References

Amol Agrawal. 2016. Clickbait detection using deep learning. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 268–272.

Leila Behboudi Angizeh and Mohammad Reza Keyvanpour. 2024. Detecting fake news using advanced language models: Bert and roberta. In *2024 10th International Conference on Web Research (ICWR)*, pages 46–52. IEEE.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16.

Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: Recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD '15, page 15–19, New York, NY, USA. Association for Computing Machinery.

Hyewon Choi, Yejun Yoon, Seunghyun Yoon, and Kunwoo Park. 2022. How does fake news use a thumbnail? CLIP-based multimodal detection on the unrepresentative news image. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 86–94, Dublin, Ireland. Association for Computational Linguistics.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Moch Fakhruroji, Cecep Suryana, and Aep Wahyudin. 2023. Clickbait journalism: Media logics in journalism practices on online media. *Communicatus: Jurnal Ilmu komunikasi*, 7(2):229–244.

Sarah Farhadpour, Timothy A Warner, and Aaron E Maxwell. 2024. Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. *Remote Sensing*, 16(3):533.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Sani Muhamad Isa, Gary Nico, and Mikhael Permana. 2022. Indobert for indonesian fake news detection. *ICIC Express Letters*, 16(3):289–297.

Dongxu Li, Junnan Li, and Steven Hoi. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Justus Mattern, Yu Qiao, Elma Kerz, Daniel Wiechmann, and Markus Strohmaier. 2021. FANG-COVID: A new large-scale benchmark dataset for fake news detection in German. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 78–91, Dominican Republic. Association for Computational Linguistics.

Logan Molyneux and Mark Coddington. 2020. Aggregation, clickbait and their effect on perceptions of journalistic credibility and quality. *Journalism Practice*, 14(4):429–446.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

BH Puneetha, Manoj Kumar, BS Prashanth, and Ajay Kumara. 2025. Enhancing fairness and performance: Scalable hybrid solutions for class imbalance in big data analytics. *Procedia Computer Science*, 258:1050–1061.

Kate Scott. 2021. You won't believe what's in this paper! clickbait, relevance and the curiosity gap. *Journal of pragmatics*, 175:53–66.

Ankit Shrestha, Audrey Flood, Saniat Sohrawardi, Matthew Wright, and Mahdi Nasrullah Al-Ameen. 2024. A first look into targeted clickbait and its countermeasures: The power of storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Andika William and Yunita Sari. 2020. Click-id: A novel dataset for indonesian clickbait headlines. *Data in brief*, 32:106231.

Jianxing Yu, Shiqi Wang, Han Yin, Zhenlong Sun, Ruobing Xie, Bo Zhang, and Yanghui Rao. 2024. Multimodal clickbait detection by de-confounding biases using causal representation inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10300–10317, Miami, Florida, USA. Association for Computational Linguistics.

A Scopus query

Listing 1: Scopus Search Query for Fake News Detection

```
TITLE-ABS-KEY("fake news detection")
```

Listing 2: Scopus Search Query for fake news dataset

```
TITLE-ABS-KEY ("datasets for fake news")
```

B LLaMA 3 prompt

```
<|im_start|>system<|im_sep|>
You are an assistant that always answers
either 'clickbait' or 'non-clickbait'
when evaluating Indonesian news articles.
Do not explain. Do not output anything
else.
<|im_end|>
<|im_start|>user<|im_sep|>
Judul: [TITLE]
Isi: [CONTENT]
Apakah ini clickbait atau non-clickbait?
<|im_end|>
<|im_start|>assistant<|im_sep|>
```

C Experimental settings

Model	Epochs	LR	K-Folds
BiLSTM	5	1e-3	5
CNN	5	1e-3	5
BERT (en)	3	2e-5	5
BERT (m)	3	2e-5	5
BERT (id)	3	2e-5	5
LLaMA 3 (8B)	–	–	–

Table 3: Settings for text-only and LLM-based models. LLaMA 3 uses zero-shot prompting via Ollama API.

Model	Epochs	LR	K-Folds
ResNet50	10	1e-5	5
EfficientNet	10	1e-5	5
CLIP	5	2e-5	5
BLIP	5	2e-5	5

Table 4: Settings for image-only models. CLIP and BLIP are used without fine-tuning.

Model	Epochs	LR	K-Folds
BiLSTM + ResNet50	10	1e-5	5
CNN + ResNet50	10	1e-5	5
BERT (en) + ResNet50	3	2e-5	5
BERT (m) + ResNet50	3	2e-5	5
BERT (id) + ResNet50	3	2e-5	5
CLIP + BERT (id)	3	2e-5	5
BLIP + BERT (id)	3	2e-5	5

Table 5: Settings for multimodal models. Blue = English BERT, Red = Multilingual BERT, Green = Indonesian BERT.

D Results

Model	Modality	Metric	Value
ResNet50	Images	Accuracy	0.4999
		Precision	0.2003
		Recall	0.4000
		F1-score	0.2669
		ROC AUC	0.4950
EfficientNet	Images	Accuracy	0.4999
		Precision	0.2003
		Recall	0.4000
		F1-score	0.2669
		ROC AUC	0.4950

Table 6: Performance metrics of ResNet50 and EfficientNet.

Model	Modality	Metric	Value
BiLSTM	Text	Accuracy	0.7543
		Recall	0.6477
		Precision	0.6700
		F1 Score	0.6574
		ROC-AUC	0.8098
CNN	Text	Accuracy	0.7582
		Recall	0.6844
		Precision	0.6645
		F1 Score	0.6732
		ROC-AUC	0.8224
BERT (en)	Text	Accuracy	0.7749
		Recall	0.6177
		Precision	0.7361
		F1 Score	0.6647
		ROC-AUC	0.8344
BERT (m)	Text	Accuracy	0.7901
		Recall	0.6736
		Precision	0.7377
		F1 Score	0.6983
		ROC-AUC	0.8561
BERT (id)	Text	Accuracy	0.8062
		Recall	0.7442
		Precision	0.7319
		F1 Score	0.7365
		ROC-AUC	0.8767

Table 7: Performance metrics of various models for fake news detection using text-based modalities. The blue-highlighted row represents an English-pretrained BERT model, the red denotes a multilingual-pretrained BERT model, and the green corresponds to an Indonesian-pretrained BERT model.

Table 8: Average performance metrics across all folds using text+images modality.

Model	Modality	Metric	Value
CNN-ResNet	Text+Images	Accuracy	0.5500
		Precision	0.6463
		Recall	0.7581
		F1 Score	0.5519
		ROC-AUC	0.6229
BiLSTM-ResNet	Text+Images	Accuracy	0.6615
		Precision	0.6534
		Recall	0.8761
		F1 Score	0.7303
		ROC-AUC	0.8003
BERT (en)	Text+Images	Accuracy	0.7653
		Precision	0.7010
		Recall	0.6673
		F1 Score	0.6727
		ROC-AUC	0.8366
BERT (m)	Text+Images	Accuracy	0.7921
		Precision	0.7375
		Recall	0.6757
		F1 Score	0.7022
		ROC-AUC	0.8545
BERT (id)	Text+Images	Accuracy	0.8017
		Precision	0.7321
		Recall	0.7287
		F1 Score	0.7279
		ROC-AUC	0.8771

Model	Modality	Accuracy	Precision	Recall	F1 Score	ROC AUC
CLIP	Image Only	0.6402	0.0000	0.0000	0.0000	0.5000
BLIP	Image Only	0.6402	0.0000	0.0000	0.0000	0.4970
CLIP + BERT (id)	Multimodal	0.8011	0.7212	0.7289	0.7251	0.8663
CLIP + BERT (id)	Multimodal	0.9377	0.9767	0.8559	0.9123	0.9813
CLIP + BERT (id)	Multimodal	0.9840	0.9861	0.9697	0.9778	0.9970
CLIP + BERT (id)	Multimodal	0.9807	0.9936	0.9524	0.9725	0.9945
CLIP + BERT (id)	Multimodal	0.9917	0.9834	0.9939	0.9886	0.9992
BLIP + BERT (id)	Multimodal	0.8127	0.7861	0.6585	0.7350	0.8487
BLIP + BERT (id)	Multimodal	0.9570	0.9200	0.9709	0.9714	0.9866
BLIP + BERT (id)	Multimodal	0.9614	0.9917	0.9015	0.9879	0.9887
BLIP + BERT (id)	Multimodal	0.9939	0.9954	0.9877	0.9946	0.9991
BLIP + BERT (id)	Multimodal	0.9950	0.9954	0.9908	0.9969	0.9978
LLaMA 3 (8B)	Text Only (LLM)	0.5409	0.6828	0.4259	0.5246	–

Table 9: Performance results of CLIP, BLIP, CLIP + BERT (id) , BLIP + BERT (id) , and LLaMA 3 (8B). While CLIP and BLIP alone fail to capture class distinction (F1 = 0), fusion with BERT (id) results in state-of-the-art performance. LLaMA 3 shows moderate performance as a zero-shot text-only model.