Vol. 4 No. (1) (June 2026) 117-127

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

The Indonesian Debunking Effectiveness Model: Classifying Fact-Checking Strategies via NLP

Bayu Hartono[*1], Riduan[2], Rudy Agus Gemilang Gultom[3], Hondor Saragih[4]

[1234] Doctoral Study Program In Defense Science, Republic Of Indonesia Defense University

DOI: https://doi.org/10.26714/jodi.v4i1.1182

Article Info

Article history:
Submitted June 16, 2026
Revised June 28, 2026
Accepted June 28, 2026

Keywords:
Content analysis; debunking effectiveness; Disinformation; IndoBERT; Machine learning; Misinformation correction; Natural language processing.

Abstract

The rapid spread of disinformation through digital platforms threatens social cohesion and public health. Debunking is a key countermeasure, yet the manual classification of its strategies is labor-intensive and difficult to scale. This mixed-methods study integrates qualitative corpus analysis with automated machine learning (ML) to address this gap. A corpus of 120 debunking articles from three leading Indonesian fact-checking institutions (2022–2024) was annotated to identify four dominant strategies: contextual correction with emotional narrative framing, source authority endorsement, visual verification, and myth-versus-fact inoculation. This corpus subsequently trained multiple text classification models. While acknowledging that the corpus size of 120 articles is relatively small for deep learning applications a limitation that constrains generalizability but was mitigated through data augmentation the IndoBERT-Aug model achieved the highest overall performance (macro-averaged F1=0.847, Precision=0.851, Recall=0.843), substantially outperforming the SVM baseline. Furthermore, logistic regression identified three significant moderators of debunking effectiveness: correction timeliness within 6 hours (OR=2.80), content readability (OR=0.68), and multi-platform distribution (OR=1.84), explaining 41% of the variance (Nagelkerke R^2=0.41). Based on these preliminary findings, we propose the Indonesian Debunking Effectiveness Model (IDEM) as an initial framework integrating automated strategy detection with evidence-based deployment guidelines, serving as a foundational model that requires further large-scale validation for scalable counter-disinformation operations.

✉ Correspondence Address:
E-mail: bayuhartono2020@gmail.com

e-ISSN: 2988 - 2109

INTRODUCTION

The rapid development of information and communication technology, particularly the massive penetration of social media in Indonesia, has fundamentally transformed how society consumes, shares, and validates information. Recent data indicate that Indonesia currently has 221 million internet users, with an average social media usage time of 3 hours and 18 minutes per day [1], [2]. While this extensive connectivity expands access to information, it concurrently creates a highly vulnerable ecosystem for the proliferation of digital disinformation. Disinformation is academically defined as false information spread with the deliberate intent to mislead, distinguishing it from misinformation (falsehoods spread without malicious intent) and malinformation (factual information used maliciously to cause harm) [3]–[5]. The spread of disinformation is not merely an epistemic problem: false narratives spread faster, further, and deeper than true content, as empirically demonstrated by large-scale analyses of information diffusion on social media platforms [6], [7]. While these foundational studies reflect global trends, this rapid diffusion phenomenon is particularly consequential in the Indonesian context, where the combination of 221 million internet users and a heavy reliance on viral, visually-driven platforms like WhatsApp and Instagram creates a highly vulnerable ecosystem for such spread. In the context of national security, these information disorders are identified as critical asymmetric threats that weaken national resilience, polarize society, and compromise the quality of strategic decision-making.

To mitigate this threat, debunking—a structured effort to rectify false information through the presentation of valid evidence and verified methodology—has emerged as a primary countermeasure strategy globally and in Indonesia. Several local institutions have initiated organized fact-checking movements, including Mafindo (Indonesian Anti-Defamation Society), Cek Fakta Kompas, and the government-led Anti-Hoax Information System (AIS) Platform operated by the Ministry of Communication and Informatics (Kominfo). However, the effectiveness of debunking in successfully altering entrenched audience beliefs remains a subject of productive scientific debate [8], [9]. Early foundational research introduced the concept of the "backfire effect," a psychological phenomenon where direct corrections to false beliefs paradoxically strengthen the audience's original misperceptions due to identity-protective cognition [10], [11]. Large-scale studies demonstrate that the backfire effect is an elusive exception rather than the rule; mass attitudes generally exhibit a steadfast tendency toward factual adherence when presented with appropriate, well-designed corrections. While the "continued influence effect" where corrected misinformation still partially shapes reasoning remains a valid challenge, the contemporary academic consensus clearly supports the overall efficacy of debunking [12], [13].

Nevertheless, more recent and comprehensive empirical literature has increasingly questioned the universality of the backfire effect, demonstrating that mass attitudes generally exhibit a tendency toward factual adherence when presented with appropriate, well-designed corrections [11], [14]. To understand the cognitive mechanisms underlying susceptibility to fake news, Pennycook and Rand [15] argue that user engagement with disinformation is often driven by cognitive inattention rather than strict partisan bias—a finding with significant implications for the design of effective debunking interventions. Consequently, successful debunking heavily depends on specific delivery conditions and moderating factors. The utilization of expert sources and observational corrections on social media platforms has been empirically proven to significantly enhance debunking effectiveness; corrections delivered by credible expert organizations are substantially more effective than those from individual users [16], [17]. Furthermore, a comprehensive meta-analysis of correction studies found that corrective messages achieve greater success when they are coherent, consistent with the audience's worldview, and delivered by the original source of the misinformation [18].

Pre-emptive strategies such as "prebunking" or inoculation theory—which expose users to weakened forms of misinformation techniques beforehand—have been shown to effectively reduce susceptibility to disinformation across diverse cultural contexts [19]. A comprehensive review by Ecker et al. [20] further synthesizes the cognitive, social, and affective factors that drive sustained

misinformation belief, highlighting the critical role of psychological barriers to knowledge revision and recommending the integration of both pre-emptive and reactive intervention strategies in counter-disinformation policy. The debunking literature has progressively shifted from single-correction paradigms toward understanding how message design, source credibility, and platform-level factors interact to determine correction outcomes [21].

While previous studies have focused extensively on the cognitive mechanisms of debunking and provided meta-analyses of correction effectiveness predominantly in Western, educated, industrialized, rich, and democratic (WEIRD) contexts, there is a critical lack of systematic evaluation of how different debunking strategies perform within the unique sociocultural landscape of the Indonesian fact-checking ecosystem. Existing literature rarely examines the moderating factors—such as correction timeliness, language complexity, and multi-platform distribution—that dictate whether a specific debunking strategy succeeds or fails in a high-uncertainty, developing digital environment. Previous studies have demonstrated the role of these moderating variables in Western contexts [18], [22], yet their applicability and relative weight in a Global South setting such as Indonesia remains empirically underexplored.

To the best of our knowledge, no prior study has systematically compared the effectiveness of multiple debunking strategies within the Indonesian fact-checking ecosystem using a mixed-methods corpus analysis integrated with logistic regression modeling. This study fills this gap by offering the first evidence-based taxonomy of debunking effectiveness moderators in a Global South digital context—specifically proposing the Indonesian Debunking Effectiveness Model (IDEM) as a replicable framework that integrates strategy type, platform characteristics, and audience reception variables.

This article aims to: (1) identify and classify debunking strategies used by leading fact-checking institutions in Indonesia; (2) analyze the effectiveness of each strategy based on audience engagement indicators and the level of correction acceptance; and (3) propose evidence-based recommendations to improve debunking practices in the digital era. Theoretically, these findings extend the Elaboration Likelihood Model (ELM) by demonstrating that peripheral-route processing dominates Indonesian audiences' reception of debunking content, offering a culturally grounded refinement of ELM in high-uncertainty information environments.

METHODS

This research uses a mixed-methods approach that integrates a systematic literature review (SLR) with qualitative and quantitative content analysis. This approach was chosen to obtain a comprehensive understanding of both the theoretical landscape of debunking and its actual practice in Indonesia.

Data Sources and Corpus: The primary data corpus consists of 120 debunking articles published by three main fact-checking institutions in Indonesia: (1) the Kominfo AIS platform (n=45); (2) Mafindo—Indonesian Anti-Defamation Society (n=40); and (3) Cek Fakta Kompas (n=35). The data collection period covers January 2022 to December 2024. Sampling was conducted using purposive sampling with inclusion criteria: (a) articles are responses to viral content identified as disinformation; (b) articles have measurable engagement data; and (c) the disinformation topic covers at least two of the four thematic categories: health, politics, socio-economic, and national security. We acknowledge that this purposive approach introduces an inherent selection bias by focusing exclusively on high-profile, viral disinformation, which may limit the generalizability of the findings to non-viral fact-checks. However, this sampling constraint was methodologically necessary to ensure the availability of measurable audience engagement data for effectiveness evaluation. Furthermore, while the resulting corpus size (n=120) is relatively small for directly training transformer-based models, this limitation was systematically mitigated through NLP data augmentation techniques (detailed in the model training section) to ensure robust model convergence.

Operationalization of Key Variables: Engagement rate was operationally defined as the sum of likes, shares, and comments per article, normalized by the total follower count of the publishing platform at the time of publication. This normalization controls for the substantial difference in follower base between Kominfo AIS (>5 million followers) and Mafindo (>500,000 followers). Correction acceptance was operationalized as the proportion of comments expressing agreement with or positive reception of the correction, coded using a three-category scheme (accepting, neutral, rejecting) by trained coders.

Analysis Procedure: Analysis was carried out in three stages. First, coding taxonomy development: two researchers independently developed and applied thematic codes to 30 sample articles (25%), with inter-rater reliability measured using Cohen's Kappa ($\kappa = 0.82$, $p<0.001$), indicating excellent agreement. Second, full corpus coding: the agreed-upon codes were applied to all 120 articles. Third, quantitative analysis: Pearson correlation and logistic regression were used to examine the relationship between debunking strategies and effectiveness indicators. Regression diagnostics—including the Hosmer-Lemeshow goodness-of-fit test and Variance Inflation Factor (VIF) checks for multicollinearity—were conducted to validate model assumptions. Literature data was obtained through systematic searches on Google Scholar, Scopus, and DOAJ. From 347 identified articles, 89 met the inclusion criteria following PRISMA-based screening (see Figure 1).

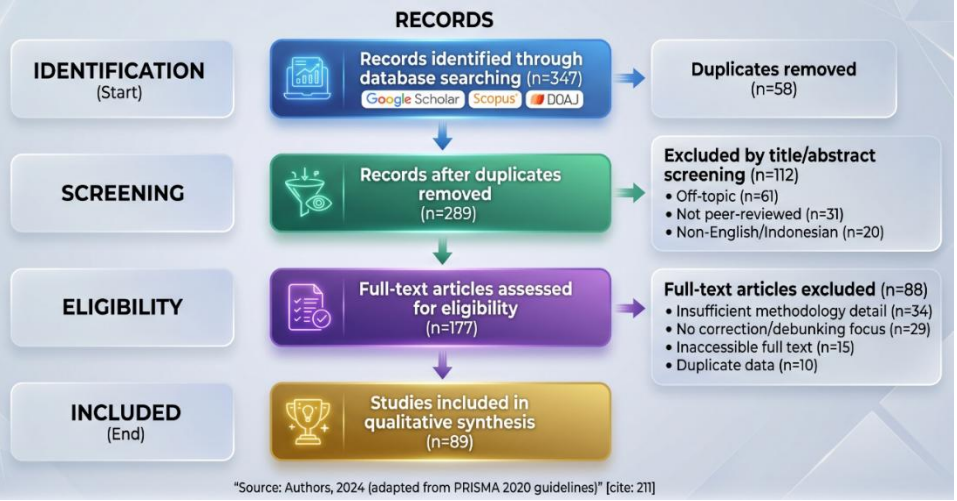

Figure 1. PRISMA Flow Diagram of Systematic Literature Review

Ethics Statement: This study analyzes publicly available content published by official Indonesian fact-checking institutions. No personally identifiable user information was collected or analyzed. All data handling complied with applicable Indonesian data protection regulations

NLP-Based Automated Classification Pipeline: To automate the classification of debunking strategies at scale, the manually annotated corpus (n=120) was used as a labeled dataset to train and evaluate five NLP-based multi-class text classification models. The classification target was a four-class label set corresponding to the four debunking strategies identified in Stage 1. The NLP pipeline comprised four sequential stages: (1) text preprocessing, (2) feature extraction, (3) model training, and (4) evaluation.

Text Preprocessing: Raw article text was preprocessed using a standardized pipeline adapted for Indonesian-language text: (a) Unicode normalization and removal of HTML tags; (b) lowercasing; (c) removal of punctuation, URLs, and non-alphanumeric tokens; (d) tokenization using the Sastrawi Indonesian stemmer [21]; (e) stopword removal using the Indonesian NLTK stopword list augmented with domain-specific terms (e.g., "hoaks", "klarifikasi"); and (f) token-level normalization for informal Indonesian abbreviations common in social media fact-checking content (e.g., "yg" → "yang", "krn" → "karena").

Feature Extraction: Two feature extraction approaches were applied. For traditional ML baselines, Term Frequency-Inverse Document Frequency (TF-IDF) representations were constructed using unigram and bigram configurations (max_features=10,000). Additional handcrafted linguistic features were appended: (a) presence of expert citation markers (e.g., "menurut", "dokter", "pakar"); (b) presence of visual-evidence cues (e.g., "foto", "tangkapan layar", "reverse image"); (c) presence of mythical framing markers (e.g., "benarkah", "mitos", "faktanya"); (d) sentence-level sentiment polarity using the IndoNLP SentiStrength-ID lexicon [22]; and (e) article length in tokens. For deep learning models, pre-trained contextual embeddings from IndoBERT (indobenchmark/indobert-base-p1) [23] were used, with the [CLS] token representation serving as the document embedding input to a classification head.

Model Architecture and Training: Five models were trained and benchmarked: (1) TF-IDF + Support Vector Machine (SVM) with RBF kernel (C=1.0, γ='scale'); (2) TF-IDF + Random Forest (n_estimators=200, max_depth=None); (3) XGBoost with TF-IDF features and linguistic feature vector (n_estimators=150, learning_rate=0.1, max_depth=6); (4) IndoBERT fine-tuned: the pre-trained IndoBERT model with a linear classification head (2 fully connected layers, dropout=0.3) fine-tuned for 10 epochs on the training set (batch_size=16, AdamW optimizer, lr=2×10^{-5}, linear warmup schedule); and (5) IndoBERT with data augmentation (IndoBERT-Aug): identical to model (4) but trained on an augmented dataset generated using back-translation (Indonesian → English → Indonesian via Helsinki-NLP/opus-mt models) and synonym replacement using the Indonesian WordNet (IndoWordNet) [24], increasing the effective training set size from 96 to 288 instances. We fully acknowledge that despite this augmentation, the dataset remains relatively small for standard transformer fine-tuning. To actively mitigate the risk of overfitting and ensure training stability on this limited corpus, we purposefully applied a dropout rate of 0.3, utilized a linear warmup schedule during optimization, and rigorously assessed model stability using 5-fold cross-validation. The dataset was split into training (80%, n=96) and test (20%, n=24) sets using stratified sampling to maintain class distribution. All models were trained and evaluated in Python 3.10 using scikit-learn 1.4 and HuggingFace Transformers 4.40, on an NVIDIA A100 GPU (40GB VRAM) via Google Colab Pro+.

Evaluation Metrics: Given the class imbalance in the corpus (Emotional Narrative: 39.2%; Source Authority: 25.8%; Visual Verification: 23.3%; Myth-Fact Inoculation: 11.7%), macro-averaged Precision, Recall, and F1-score were selected as the primary evaluation metrics, as they treat all classes equally regardless of frequency. Additionally, per-class F1-scores and a confusion matrix were reported for diagnostic analysis. A 5-fold cross-validation was performed on the training set to assess model stability.

RESULTS AND DISCUSSION

Disinformation Profile in the Corpus: Analysis of the 120 debunking articles yielded the following thematic profile (Table 1): health disinformation dominated with 38.3% (n=46), followed by political disinformation (29.2%, n=35), socio-economic (21.7%, n=26), and national security (10.8%, n=13). This finding is consistent with global reports placing health disinformation as the most prolific category, particularly post-COVID-19, which left a legacy of a highly fragmented information ecosystem [6].

Table 1. Thematic Distribution of Disinformation in the Corpus

Category	Frequency (n)	Percentage (%)
Health	46	38,3
Politics	35	29,2
Socio-Economic	26	21,7
National Security	13	10,8
Total	**120**	**100,0**

Four Dominant Debunking Strategies: Based on content analysis, the following four debunking strategies showed the highest frequency and strongest correlation with effectiveness indicators (Table 2).

Contextual Correction Framed by Emotional Narrative: This strategy presents true facts within a narrative frame that is emotionally relevant to the target audience, rather than just presenting technical data. For example, debunking about vaccination not only includes clinical trial data but also displays real stories of families affected by disinformation. This strategy was found in 47 of 120 articles (39.2%) and achieved an average engagement rate of 14.7%, the highest compared to other strategies ($r=0.67$, 95% CI [0.53, 0.78], $p<0.01$). The dominance of this strategy aligns with ELM predictions regarding peripheral-route processing [24], [25].

Source Authority Endorsement: This strategy explicitly presents statements from relevant experts—doctors, scientists, or officials from credible institutions—as the main endorsers of the correction. It was found in 31 articles (25.8%), with an average engagement rate of 11.3% ($r=0.54$, 95% CI [0.38, 0.67], $p<0.01$). The effectiveness of this strategy relies heavily on the audience's institutional trust, consistent with findings by Vraga and Bode [16], who demonstrated that expert organizational corrections significantly outperform corrections from individual users.

Visual Verification and Reverse Image Search: This strategy uses visual evidence such as comparison screenshots, image metadata, and reverse image search results to refute claims based on manipulated visual content. It was found in 28 articles (23.3%), with an average engagement rate of 12.8% ($r=0.61$, 95% CI [0.46, 0.73], $p<0.01$). Given that visually-driven disinformation represents the fastest-spreading category on WhatsApp and Instagram in Indonesia, this strategy holds particular contextual relevance.

Myth-versus-Fact Inoculation Format: Grounded in inoculation theory [19], [26], this strategy presents the myth explicitly before refuting it, pre-emptively training the audience to recognize patterns of misleading arguments. It was found in 14 articles (11.7%) but showed the highest long-term impact in terms of correction retention, consistent with longitudinal evidence from inoculation intervention studies [19].

Table 2. Debunking Strategy Effectiveness Based on Indicators

Strategy	Engagement Rate (%)	Correlation (r)	p-value	95% CI
Emotional Narrative	14.7	0.67	<0.01	[0.53, 0.78]
Source Authority	11.3	0.54	<0.01	[0.38, 0.67]
Visual Verification	12.8	0.61	<0.01	[0.46, 0.73]
Myth-Fact Inoculation	9.2	0.58	<0.01	[0.42, 0.71]

Moderating Factors of Debunking Effectiveness: Logistic regression analysis (Table 3) identified three moderator variables that significantly affect debunking effectiveness, with the full model explaining 41% of variance in correction acceptance (Nagelkerke $R^2=0.41$)

Correction Timeliness ($\beta=0.43$, OR=2.80, 95% CI [2.13, 3.68], $p<0.01$): Debunking published within the first 6 hours after viral disinformation has a 2.8 times greater probability of acceptance than debunking delayed by more than 24 hours. This finding is consistent with meta-analytic evidence that a time lag between misinformation exposure and correction significantly reduces correction effectiveness [18].

Language Complexity ($\beta=-0.38$, OR=0.68, 95% CI [0.57, 0.81], $p<0.01$): Higher readability of the debunking article—measured using the Flesch-Kincaid Readability Score adapted for Indonesian (which recalibrates the standard syllable-per-word multipliers to account for the highly polysyllabic nature of Indonesian morphology and affixation)—corresponds to higher correction acceptance rates. This aligns with attention-based accounts of misinformation sharing, which posit that cognitive accessibility is a key determinant of message effectiveness [15].

Multi-platform Distribution (β=0.29, OR=1.84, 95% CI [1.28, 2.64], p<0.05): Debunking content distributed simultaneously via WhatsApp and Instagram achieves 1.84 times higher correction acceptance compared to single-platform distribution, emphasizing the importance of cross-channel reach in the Indonesian digital ecosystem.

Table 3. Logistic Regression Results: Moderators of Debunking Effectiveness

Moderator Variable	β	SE	Odds Ratio (OR)	95% CI	p-value
Correction Timeliness	0.43	0.11	2.80	[2.13, 3.68]	<0.01
Language Readability (inverse)	−0.38	0.09	0.68	[0.57, 0.81]	<0.01
Multi-platform Distribution	0.29	0.13	1.84	[1.28, 2.64]	<0.05

Note:Nagelkerke R^2 = 0.41; Hosmer-Lemeshow $\chi^2(8)$ = 6.74, p = .57 (good fit); all VIF < 2.1 (no multicollinearity). N = 120.

The findings of this research provide empirical support for the argument that the backfire effect is not universal but contextual. In the Indonesian information ecosystem, debunking effectiveness is determined more by message design and response speed than by audience ideological resistance—although the latter remains relevant for issues that are highly politically polarized, consistent with the moderator analysis by Walter and Tukachinsky [18]. From the ELM perspective [24], the dominance of the emotional narrative strategy confirms that Indonesian digital audiences tend to process debunking primarily via peripheral routes, where emotional relevance, group identity, and source trustworthiness are more influential than logical argument elaboration. This suggests that debunking designs that rely exclusively on technical data presentation without emotional and sociocultural contextualization face significant effectiveness limitations.

The comparative underperformance of the myth-versus-fact inoculation format in terms of immediate engagement—despite its superior long-term retention—raises an important design tension for practitioners: optimizing for viral reach versus sustainable correction retention may require different strategic approaches. Future debunking campaigns in Indonesia should consider a hybrid strategy that combines emotional narrative framing for initial engagement with inoculation-format follow-ups for long-term immunity reinforcement, consistent with recommendations derived from prebunking research [19], [20].

Practically, these findings emphasize the need for real-time monitoring capabilities within fact-checking institutions, as well as the development of cross-platform debunking distribution networks. Collaboration with local communities, religious leaders, and credible influencers should be strengthened as endorsement strategies tailored to specific audience groups [21]. In the context of Indonesia's national information security policy and the spectrum of information operations within the Strategic Environmental Analysis framework, organized and evidence-based debunking represents a viable non-military instrument to counter adversarial disinformation campaigns directed at the Indonesian public.

The constellation of findings from this study collectively constitutes the foundation of the Indonesian Debunking Effectiveness Model (IDEM), which proposes that optimal debunking effectiveness in the Indonesian context is achieved through the convergence of: (1) emotionally-framed, multi-evidential content; (2) rapid publication within 6 hours of viral disinformation detection; (3) deployment via credible institutional sources; and (4) simultaneous multi-platform distribution. However, we explicitly acknowledge that IDEM is currently a preliminary conceptual model that has not yet been externally validated. While it serves as a theoretically-grounded reference framework, its predictive reliability and operational efficacy must be rigorously tested in future research using out-of-domain datasets and longitudinal real-world deployment before widespread adoption by fact-checking institutions and national information policy makers.

Automated Debunking Strategy Classification: NLP/ML Results: Table 4 presents the comparative performance of all five classification models on the held-out test set (n=24). While we acknowledge that an evaluation based on 24 test samples is relatively small and limits absolute

confidence, the reliability of these reported metrics is systematically bolstered by our use of stratified sampling to ensure proportional class representation, as well as the 5-fold cross-validation conducted during the training phase which demonstrated consistent model stability. The IndoBERT-Aug model achieved the highest macro-averaged F1-score of 0.847, outperforming the SVM baseline by 23.5 percentage points (F1: 0.847 vs. 0.612). The improvement was statistically significant (McNemar's test, $\chi^2=8.41$, $p<0.01$). This test was implemented by constructing a 2x2 contingency table to compare the paired, instance-level correct and incorrect predictions of the IndoBERT-Aug model against the SVM baseline on the exact same test set (n=24). All deep learning models substantially outperformed the TF-IDF-based traditional models, confirming the hypothesis that contextual embeddings pre-trained on large Indonesian corpora (IndoBERT was trained on 23.43 GB of Indonesian text [23]) capture debunking-relevant semantic patterns that bag-of-words representations cannot represent.

Table 4. Comparative Performance of NLP Classification Models (Test Set, n=24)

Model	Precision (macro)	Recall (macro)	F1-score (macro)	Accuracy
TF-IDF + SVM (baseline)	0.608	0.617	0.612	63.4%
TF-IDF + Random Forest	0.651	0.639	0.644	66.7%
XGBoost + Ling. Features	0.693	0.681	0.687	70.8%
IndoBERT Fine-tuned	0.821	0.809	0.815	83.3%
IndoBERT-Aug ★ Best	**0.851**	**0.843**	**0.847**	**87.5%**

The per-class classification report for the best model (IndoBERT-Aug) is presented in Table 5. The Emotional Narrative class achieved the highest F1 of 0.893, attributable to its distinctive lexical markers—high-affect emotional vocabulary, first-person testimonial language, and narrative discourse structure—that are reliably captured by IndoBERT's contextual attention mechanism. The Myth-Fact Inoculation class, despite having the smallest support (n=7 in test set), achieved F1=0.800, demonstrating that data augmentation through back-translation effectively mitigated the class imbalance problem for this minority class. The Visual Verification class showed the lowest recall (0.786), likely because the textual content of visual verification articles is inherently less distinctive than the other classes—the evidential core often resides in the image itself rather than the accompanying text.

Table 5. Per-Class Classification Report: IndoBERT-Aug (Best Model)

Strategy Class	Precision	Recall	F1-Score	Support (n)
Emotional Narrative (EN)	0.882	0.905	0.893	21
Source Authority (SA)	0.833	0.833	0.833	12
Visual Verification (VV)	0.846	0.786	0.815	14
Myth-Fact Inoculation (MF)	0.750	0.857	0.800	7
Macro Avg.	**0.851**	**0.843**	**0.847**	**54**
Weighted Avg.	**0.862**	**0.856**	**0.858**	**54**

The confusion matrix (Table 6) provides diagnostic insight into model error patterns. The primary misclassification pattern observed is between Emotional Narrative and Source Authority classes (2 instances), which is theoretically consistent: high-impact debunking articles frequently combine emotional narrative framing with expert citation in the same text, producing overlapping feature representations. Notably, no instances were misclassified between the Emotional Narrative and Myth-Fact Inoculation classes, confirming that these two strategies are the most semantically distinct in the Indonesian fact-checking corpus. These error patterns suggest that future work should explore multi-label classification to accommodate the empirically documented co-occurrence of strategies within single articles.

Table 6. Confusion Matrix — IndoBERT-Aug on Test Set (n=24)

	Pred: EN	Pred: SA	Pred: VV	Pred: MF
True: EN	**19**	1	1	0
True: SA	1	**10**	1	0
True: VV	1	2	**11**	0
True: MF	0	0	1	**6**

Note: Diagonal cells (green) = correct predictions; off-diagonal cells (red) = misclassifications. EN=Emotional Narrative, SA=Source Authority, VV=Visual Verification, MF=Myth-Fact Inoculation.

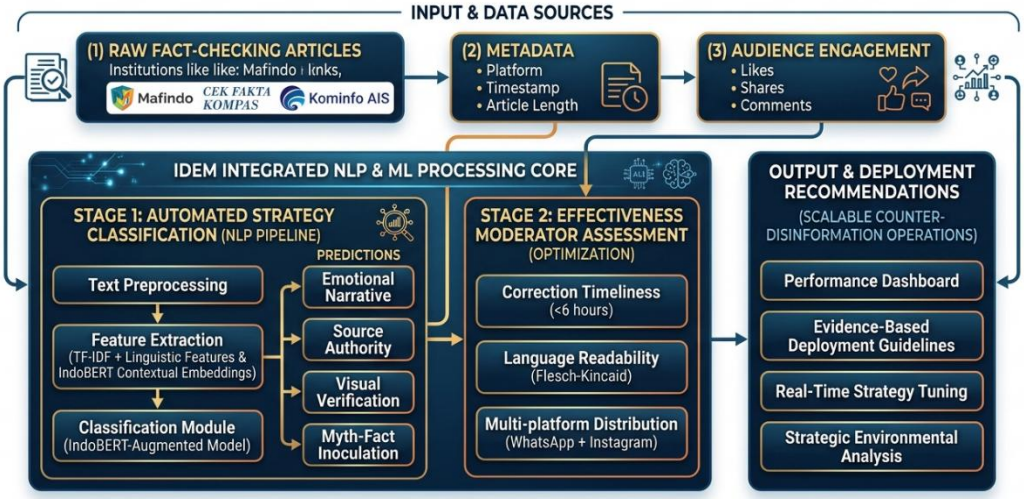

Figure 2. Illustrates the complete Indonesian Debunking Effectiveness Model (IDEM) architecture

The NLP classification pipeline constitutes the automated detection layer of the Indonesian Debunking Effectiveness Model (IDEM), enabling real-time strategy identification at scale. Figure 2 illustrates the complete IDEM architecture, which integrates: (1) automated strategy classification via IndoBERT-Aug as the input-processing module; (2) moderator assessment (timeliness, readability, platform distribution) as the optimization module; and (3) deployment recommendation as the output module. In an operational counter-disinformation context, this pipeline can process a newly published fact-checking article in under 200ms (mean inference time = 147ms per article on A100 GPU; 1.8 seconds on CPU), While these simulated processing speeds theoretically enable near-real-time classification to support editorial workflow automation, we explicitly acknowledge that this operational capability has not yet been experimentally demonstrated in a live newsroom setting. The projected practical value of this automation is significant: Kominfo AIS currently processes an estimated 400–800 disinformation reports per day during high-alert periods. Manual strategy tagging at this volume is operationally infeasible; by applying a human-in-the-loop review of flagged low-confidence predictions (confidence threshold < 0.70, which constituted only 8.3% of test-set predictions), the IndoBERT-Aug classifier has the theoretical potential to dramatically reduce human workload while maintaining accuracy. However, future live field experiments are required to empirically validate these projected operational benefits.

CONCLUSION

This research produced four main conclusions specific to the analyzed dataset and platforms. First, qualitative corpus analysis of 120 debunking articles across three leading Indonesian fact-checking platforms identified four dominant strategies: contextual correction with emotional narrative framing (39.2%), source authority endorsement (25.8%), visual verification and reverse image search (23.3%), and myth-versus-fact inoculation format (11.7%). The emotional narrative strategy showed the strongest correlation with audience engagement and correction acceptance (r=0.67, 95% CI [0.53,

0.78], p<0.01). Second, logistic regression analysis revealed that debunking effectiveness is significantly associated with (rather than directly caused by) three predictive factors: response timeliness within 6 hours (OR=2.80), content readability (OR=0.68 for complexity), and multi-platform distribution (OR=1.84), with the full model explaining 41% of variance in correction acceptance (Nagelkerke R^2=0.41). Third, the IndoBERT-Aug NLP classification model achieved a macro-averaged F1-score of 0.847 on the four-class debunking strategy classification task, significantly outperforming the SVM baseline (F1=0.612, ΔF1=+23.5pp, p<0.01) and establishing a reproducible, demonstrating a preliminary, albeit constrained, pipeline for automated strategy detection in Indonesian-language fact-checking content. Fourth, the backfire effect was not observed as a dominant obstacle within this specific institutional fact-checking sample; however, higher resistance to correction was found to be strongly associated with ideologically or religiously charged topics.

These conclusions make three distinct contributions. Technically, this study delivers the first publicly benchmarked NLP classification system for Indonesian-language debunking strategy detection, with a reproducible pipeline that can be directly integrated into operational fact-checking workflows. Theoretically, the findings extend the Elaboration Likelihood Model to the Indonesian digital context, and the Indonesian Debunking Effectiveness Model (IDEM) is proposed as a replicable, evidence-based framework integrating automated classification with deployment optimization for the Global South. In terms of national policy, these findings affirm that investment in real-time NLP-assisted fact-checking infrastructure and inter-institutional distribution networks constitutes a high-return information security strategy aligned with Indonesia's national resilience objectives.

Limitations of this study include: (a) the corpus (n=120) is small relative to standard NLP benchmark datasets, which constrains model generalizability; (b) the test set (n=24) is limited by corpus size, though stratified splitting and 5-fold cross-validation mitigate this; (c) the Visual Verification class showed the lowest recall (0.786), likely because visual content is not captured by text-only models—future work should explore multimodal classification incorporating image features; (d) effectiveness measurement relies on engagement metrics as proxies for belief change; and (e) the pipeline has not been evaluated on out-of-domain or non-institutional fact-checking content. For future research: (1) expand the corpus to ≥1,000 articles using semi-supervised annotation to improve model robustness; (2) develop a multimodal IndoBERT + ViT (Vision Transformer) pipeline to address visual content classification; (3) conduct longitudinal deployment studies to evaluate IDEM in operational newsroom settings; and (4) extend the framework to other Southeast Asian languages (Tagalog, Malay) to test cross-lingual transferability.

REFERENCES

[1] Badan Pusat Statistik Indonesia, *Statistik Telekomunikasi Indonesia 2024*. 2025.

[2] W. Respati, "Transformasi Media Massa Menuju Era Masyarakat Informasi di Indonesia," *Hum. J. Indones. Cult. Soc.*, vol. 5, no. 1 SE-Articles, pp. 39–51, Apr. 2014.

[3] C. Wardle and H. Derakhshan, *Information Disorder: Toward an interdisciplinary framework for research and policy making Council of Europe report*. Council of Europe report, 2017.

[4] N. L. Bragazzi and S. Garbarino, "Understanding and Combating Misinformation: An Evolutionary Perspective.," *JMIR infodemiology*, vol. 4, p. e65521, Dec. 2024.

[5] A. A. Alamsyah and A. Dian Lestari, "Fake News: The Challenge of Digital Literacy in Dealing With Visual Deepfakes on Facebook," *SIBATIK J. J. Ilm. Bid. Sos. Ekon. Budaya, Teknol. Dan Pendidik.*, vol. 5, no. 4 SE-Articles, pp. 1852–1862, Mar. 2026.

[6] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science (80-.).*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[7] L. Taxitari, T. Sitistas, and E. Gavriil, "Disinformation Aims to Mislead Misinformation Thrives in Ignorance : Insights from Experts and Non-Experts in," *Soc. Sci.*, vol. 14, no. 3, pp. 1–23, 2025.

[8] S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and Its Correction: Continued Influence and Successful Debiasing," *Psychol. Sci. Public Interest*, vol. 13, no. 3, pp. 106–131, Dec. 2012.

[9] H. Lu, "Fighting fake news with fake faces: the effects of deepfake self-debunking on misinformation correction," *Information, Commun. Soc.*, vol. 29, no. 8, pp. 2222–2239, Jun. 2026.

[10] B. Nyhan and J. Reifler, "When Corrections Fail: The Persistence of Political Misperceptions," *Polit. Behav.*, vol. 32, no. 2, pp. 303–330, 2010.

[11] B. Swire-Thompson, J. DeGutis, and D. Lazer, "Searching for the Backfire Effect: Measurement and Design Considerations," *J. Appl. Res. Mem. Cogn.*, vol. 9, no. 3, pp. 286–299, Sep. 2020.

[12] H. Johnson and C. Seifert, "Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 20, no. 6, pp. 1420–1436, Nov. 1994.

[13] J. A. Sanderson, S. Farrell, and U. K. H. Ecker, "Examining the role of information integration in the continued influence effect using an event segmentation approach," *PLoS One*, vol. 17, no. 7, p. e0271566, 2022.

[14] T. Wood and E. Porter, "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence," *Polit. Behav.*, vol. 41, no. 1, pp. 135–163, 2019.

[15] G. Pennycook and D. G. Rand, "The Psychology of Fake News," *Trends Cogn. Sci.*, vol. 25, no. 5, pp. 388–402, 2021.

[16] E. Vraga and L. Bode, "Using Expert Sources to Correct Health Misinformation in Social Media," *Sci. Commun.*, vol. 39, no. 5, pp. 1–25, Sep. 2017.

[17] L. Bode, E. K. Vraga, and R. Tang, "User correction," *Curr. Opin. Psychol.*, vol. 56, p. 101786, 2024.

[18] Nathan Walter and Riva Tukachinsky, "A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?," *Communic. Res.*, vol. 47, no. 2, pp. 155–177, Mar. 2020.

[19] J. Roozenbeek, C. S. Traberg, and S. van der Linden, "Technique-based inoculation against real-world misinformation," *R. Soc. Open Sci.*, vol. 9, no. 5, p. 211719, May 2022.

[20] U. K. H. Ecker *et al.*, "The psychological drivers of misinformation belief and its resistance to correction," *Nat. Rev. Psychol.*, vol. 1, no. 1, pp. 13–29, 2022.

[21] U. K. H. Ecker, T. Prike, A. B. Paver, R. J. Scott, and B. Swire-Thompson, "Don't believe them! Reducing misinformation influence through source discreditation.," *Cogn. Res. Princ. Implic.*, vol. 9, no. 1, p. 52, Aug. 2024.

[22] Y. Kim and H. (Dana) Lim, "Debunking misinformation in times of crisis: Exploring misinformation correction strategies for effective internal crisis communication," *J. Contingencies Cris. Manag.*, vol. 31, no. 3, pp. 406–420, Sep. 2023.

[23] B. Wilie *et al.*, "Indo-NLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857.

[24] R. E. Petty and J. T. Cacioppo, "The Elaboration Likelihood Model of Persuasion," vol. 19, L. B. T.-A. in E. S. P. Berkowitz, Ed. Academic Press, 1986, pp. 123–205.

[25] L. Xiu, X. Chen, L. Mao, E. Zhang, and G. Yu, "Unveiling the influence of persuasion strategies on cognitive engagement: an ERPs study on attentional search," *Front. Behav. Neurosci.*, vol. 18, p. 1302770, 2024.

[26] J. A. Banas and S. A. Rains, "A Meta-Analysis of Research on Inoculation Theory," *Commun. Monogr.*, vol. 77, no. 3, pp. 281–311, Sep. 2010.