Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 5, No. 3, September 2026, pp. 585–593
e-ISSN: 2964-2922, p-ISSN: 2963-6191

Open access

Original Research

Interpreting Text-Enriched Dual-Head Multitask Learning for Indonesian Hateful Meme Detection Using Explainable AI

Selamet Riadi [a*], Emi Suryadi [b], Muhamad Masjun Efendi [c], Bahtiar Imran [d], Muhammad Zamroni Uska [e]

[a,d] Computer Systems Engineering, Faculty of Computer Science and Artificial Intelligence, Universitas Teknologi Mataram, Mataram, Indonesia
[b] Information Technology, Faculty of Computer Science and Artificial Intelligence, Universitas Teknologi Mataram, Mataram, Indonesia
[c] Information Systems, Faculty of Computer Science and Artificial Intelligence, Universitas Teknologi Mataram, Mataram, Indonesia
[e] Informatics Education, Faculty of Mathematics and Natural Sciences, Universitas Hamzanwadi, Selong, Indonesia

*Corresponding Author: didiriadijumantoro@gmail.com

ABSTRACT

Internet memes in Indonesia are frequently weaponized to disseminate implicit hate speech through sarcasm and cultural nuances. Automatically detecting such content is computationally challenging, and existing deep learning frameworks predominantly operate as opaque black boxes, lacking decision transparency. This study implements and optimizes a text-enriched dual-head multitask learning architecture utilizing IndoBERTweet to concurrently classify hatefulness and appropriateness within the INDOMEME dataset. Rather than processing raw image pixels, we employ a text-enrichment strategy where visual semantics are transcribed into textual descriptors via Optical Character Recognition and vision-language captioning. To bridge the interpretability gap, we deploy Local Interpretable Model-agnostic Explanations (LIME) to decode the internal feature attributions of the architecture. Furthermore, advanced training optimizations, encompassing cosine annealing, gradient accumulation, class-weighted loss, and dynamic threshold calibration, were engineered to enhance model generalization. Experimental evaluations demonstrate that the optimized model achieves a Macro-F1 score of 0.812 for hatefulness and 0.820 for appropriateness, surpassing the established baseline. Crucially, the LIME analysis unveils a pivotal finding: despite sharing an identical textual backbone, the hate-specific head predominantly focuses on lexicons carrying social agitation, whereas the appropriateness head prioritizes general norm violations. These empirical findings substantiate that multitask learning enriches semantic representation quality, offering a transparent framework for trustworthy content moderation.

How to Cite:
S. Riadi, E. Suryadi, M. M. Efendi, B. Imran, and M. Z. Uska, "Interpreting Text-Enriched Dual-Head Multitask Learning for Indonesian Hateful Meme Detection Using Explainable AI," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 5, no. 3, pp. 585–593, 2026.
Doi: https://doi.org/10.69916/jkbti.v5i3.576

KEYWORDS

Hateful Memes
Hate Speech Detection
Multitask Learning
Explainable AI
IndoBERTweet

ARTICLE INFO

Received: July 6, 2026
Revised: August 19, 2026
Available Online: September 1, 2026

1. INTRODUCTION

Internet memes have evolved from simple digital humor into a powerful medium for sociopolitical expression in Indonesia. On platforms like Facebook and Instagram, memes frequently blend visual satire with embedded

text to convey complex, often implicit messages. This multimodal format is increasingly exploited to propagate hate speech through sarcasm, cultural nuance, and code-mixing, making it exceptionally difficult for traditional text-only moderation systems to detect [1]. The challenge is further amplified by the low-resource nature of the Indonesian language, where meaning heavily depends on contextual interplay between image and text [2].

To address this, recent studies have proposed benchmark datasets and advanced architectures. Most notably, Pamungkas et al. introduced the INDOMEME dataset, the first expert-annotated multimodal corpus for Indonesian hateful memes, and demonstrated that a dual-head multitask learning framework could jointly optimize hatefulness and appropriateness classification [1]. Their work leveraged IndoBERTweet [3], a domain-specific transformer model proven effective for noisy Indonesian social media text. This approach builds upon foundational transformer architectures [4] and aligns with broader findings that hard-parameter sharing in multitask learning improves generalization by learning shared semantic representations across related tasks [5,6].

Despite these advances, a critical limitation persists: existing models operate as black boxes. While they achieve high accuracy, they offer no transparency into why a meme is classified as hateful. In sensitive domains like hate speech detection, interpretability is not optional—it is essential for trust, accountability, and bias mitigation [7, 8]. Although XAI techniques like LIME have been applied to single-task hate speech models [7], there remains a distinct lack of research applying XAI to decode the internal mechanics of dual-head multitask architectures, especially in the Indonesian context [9, 10].

Furthermore, while full multimodal fusion models (e.g., combining vision and language encoders) yield strong performance [11], they are computationally expensive and often impractical for real-world deployment. An efficient alternative is the text-enrichment strategy, where visual content is transcribed into textual descriptors (via OCR and AI-generated captions) and fed into a text-only model [12]. However, this approach lacks rigorous interpretability analysis to validate whether the model truly understands contextual malice or merely overfits to surface-level keywords.

To bridge these gaps, this study makes two key contributions. First, we implement and optimize a text-enriched dual-head multitask model using advanced training strategies including class-weighted loss [13], cosine annealing [14], and dynamic threshold calibration to achieve state-of-the-art performance on the INDOMEME dataset. Second, and more importantly, we apply Local Interpretable Model-agnostic Explanations (LIME) to conduct the first known interpretability analysis of a dual-head architecture for Indonesian hateful memes [15]. By comparing feature attributions between the "hate" and "appropriate" heads, we empirically demonstrate how the shared backbone disentangles linguistic cues for two distinct sociolinguistic tasks [16].

The primary objective of this research is to provide a transparent, accurate, and computationally efficient framework for Indonesian meme moderation. The remainder of this paper is structured as follows: Section 2 details the research methods; Section 3 presents experimental results and interpretability analysis; and Section 4 concludes with implications and future work.

2. RESEARCH METHODS

2.1. Dataset and Text Enrichment Strategy

This study utilizes the INDOMEME dataset, a comprehensive corpus comprising 5,023 expert-annotated Indonesian memes [1]. Each instance is annotated with dual labels: Hatefulness (Hate/Not Hate) and Appropriateness (Appropriate/Inappropriate). Although the dataset is inherently multimodal, the computational framework implemented in this study focuses exclusively on the textual modality through a text-enrichment strategy. To capture the semantic richness of the visual content without employing a dedicated, computationally heavy vision encoder, the visual information of each meme is first transcribed into textual form. Optical Character Recognition (OCR) text is extracted via Qwen2-VL-2B, and machine-generated image captions are produced by Gemini 2.5 Flash [17]. These machine-generated textual descriptors are then concatenated with topical focus annotations to form a unified textual input sequence formatted as: `[Topic] [SEP] [Caption] [SEP] [OCR]`. The dataset is partitioned into training (80%) and testing (20%) sets using stratified sampling to maintain the distribution of both classes.

2.2. Proposed Method: Dual-Head Multitask Architecture

We employ IndoBERTweet [3] as the foundational text encoder due to its robust domain adaptation to noisy Indonesian social media text. The architecture features a shared representation layer enhanced with Layer Normalization (LayerNorm) to stabilize gradient flow and mitigate the vanishing gradient problem in deep

layers. This shared layer is followed by two distinct classification heads: one for hatefulness and one for appropriateness, following recent advancements in transformer-based multitask learning for multimodal hate detection [16]. The joint multitask loss function is mathematically formulated as:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{hate}} + (1 - \lambda)\mathcal{L}_{\text{appropriate}} \tag{1}$$

Table 1. The Mathematical Notation of Multitask Loss Function

Notation	Description
$\mathcal{L}_{\text{total}}$	Total joint multitask loss optimized during the training process.
λ	Task-weighting hyperparameter used to balance relative contributions ($0 \leq \lambda \leq 1$).
$\mathcal{L}_{\text{hate}}$	Cross-entropy loss for the hatefulness detection task (Hate vs. Not Hate).
$\mathcal{L}_{\text{appropriate}}$	Cross-entropy loss for the appropriateness classification task (Appropriate vs. Inappropriate).

2.3. Advanced Training Optimization

To maximize model generalization and overcome the inherent class imbalance in the dataset, we implemented several advanced training strategies. First, Label Smoothing (ϵ) was applied to prevent the model from becoming overconfident in its predictions, thereby improving calibration and generalization in text classification tasks [18]. Second, a Class-Weighted Loss function was calculated dynamically from the training distribution to penalize misclassifications of the minority class more heavily [13]. Third, we utilized Cosine Annealing for the learning rate schedule [14], which allows for smoother convergence and helps the model escape local minima. Finally, Gradient Accumulation was employed to simulate a larger effective batch size, stabilizing the gradient updates without exceeding GPU memory limits.

2.4. Explainable AI Formulation

To interpret the dual-head architecture, we utilize LIME, a post-hoc interpretability method that explains individual predictions by approximating them locally with an interpretable model [7]. Recent unified frameworks for evaluating explainability in language models further validate the use of such feature attribution methods to ensure faithfulness in NLP classifications [15]. The explanation for a given meme instance is obtained by solving the following optimization problem:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2}$$

Table 2. The Mathematical Notation Of LIME Framework

Notation	Description
x	The original meme instance (input text) to be explained.
x'	The binary interpretable representation of x (presence/absence of words).
z'	A perturbed sample generated around x' and its binary interpretable representation.
Z'	The set of all perturbed binary samples around x'.
f	The complex black-box model (dual-head multitask classifier) being explained.
g	The interpretable surrogate model (linear model) that approximates f locally.
G	The family of interpretable models (e.g., linear models).
$\pi_x(z)$	The proximity kernel measuring the distance between x and the perturbed sample z.
$\Omega(g)$	The complexity penalty of the surrogate model g to ensure simplicity (e.g., L1 regularization).

By applying this formulation independently to the "Hate" head and the "Appropriate" head for the exact same input text, we can map the specific linguistic features that drive each task's decision.

2.5. Evaluation Metrics

Following standard practices for imbalanced classification tasks, we report Macro-Precision, Macro-Recall, and Macro-F1. Macro-averaged metrics are prioritized to ensure equal weighting of minority (Hate) and majority (Not Hate) classes, providing a more realistic assessment of model performance in real-world moderation scenarios.

3. RESULTS AND DISCUSSION

3.1. Quantitative Performance Evaluation

The implemented text-enriched dual-head multitask model was evaluated using the complete textual configuration (`Topic + Caption + OCR`). To provide a comprehensive assessment, the performance evaluation is divided into two parts: a primary comparison of the Macro-F1 scores against the established baseline, and a detailed breakdown of the internal classification metrics of the proposed model.

Table 3 presents the comparative analysis of the Macro-F1 scores for both the hatefulness detection and appropriateness classification tasks between the baseline text-enriched multitask model reported by Pamungkas et al. [1] and our advanced optimized model.

Table 3. Macro-F1 Score Comparison for Hatefulness and Appropriateness Classification

Task	Baseline (Pamungkas et al.) [1]	Proposed Model
Hatefulness Detection (Macro-F1)	0.8110	0.8120
Appropriateness Classification (Macro-F1)	0.8150	0.8197

As demonstrated in Table 3, the proposed advanced optimizations yielded noticeable improvements over the baseline. The Macro-F1 score for hatefulness detection increased marginally from 0.8110 to 0.8120. More importantly, the Macro-F1 for appropriateness classification showed a highly significant enhancement, rising from 0.8150 to 0.8197. This substantial improvement confirms that the integration of advanced training regularization effectively captures complex semantic cues in the enriched textual input.

To provide deeper visibility into the model's classification behavior, Table 4 details the comprehensive performance metrics, including Macro-Precision, Macro-Recall, and Overall Accuracy, specifically for the proposed model.

Table 4. Detailed Performance Metrics of the Proposed Text-Enriched Model

Task	Metric	Proposed Model
Hatefulness Detection	Macro-Precision	0.8086
	Macro-Recall	0.8161
	Accuracy	0.8279
Appropriateness Classification	Macro-Precision	0.8183
	Macro-Recall	0.8222
	Accuracy	0.8219

As detailed in Table 4, the proposed model achieved robust and balanced performance across all evaluation dimensions. For the hatefulness detection task, the model recorded a Macro-Precision of 0.8086 and a Macro-Recall of 0.8161, culminating in an overall accuracy of 0.8279. Meanwhile, for the appropriateness classification task, the model attained a Macro-Precision of 0.8183, a Macro-Recall of 0.8222, and an overall accuracy of 0.8219.

Notably, the Macro-Recall for the appropriateness task reached an impressive 0.8222. In the context of content moderation, a high recall is paramount to ensure that the majority of inappropriate content is successfully flagged. This exceptional recall metric is primarily attributed to the implementation of a class-weighted loss function, which effectively mitigated the inherent class imbalance in the training data. Furthermore, the optimal threshold calibration—adjusting the decision boundary from the default 0.5 to 0.38 for appropriateness classification—substantially increased the model's sensitivity to the minority "Inappropriate" class without severely compromising precision. These results confirm that the text-enrichment strategy, coupled with dynamic thresholding, produces a highly accurate, reliable, and well-calibrated moderation framework.

3.2. Impact of Text-Enrichment and Advanced Optimizations

The marginal yet consistent improvement in the Hate Macro-F1 (from 0.811 to 0.812) can be attributed to the synergistic effect of cosine annealing and label smoothing. In the context of hate speech detection, the linguistic markers are often subtle and embedded within sarcastic contexts. The label smoothing technique prevented the model from assigning extreme probability values to the training samples, thereby forcing it to learn more robust and generalized feature representations rather than memorizing the training data [18]. Furthermore, the

text-enrichment strategy proved highly effective. By converting visual semantics into structured text via OCR and captioning, the IndoBERTweet model could capture both implicit and explicit semantic cues without the computational overhead of processing raw image pixels [17]. This approach aligns with recent observations in deep learning generalization, which suggest that advanced regularization and optimization techniques can yield meaningful performance gains even on top of strong baseline architectures [19].

3.3. Qualitative Analysis: Decoding the Dual-Head Architecture

To understand the internal mechanics of the model, we applied LIME to analyze the feature attributions of both classification heads on the exact same textual input. This dual-analysis provides unprecedented visibility into how the multitask architecture processes overlapping sociolinguistic concepts.

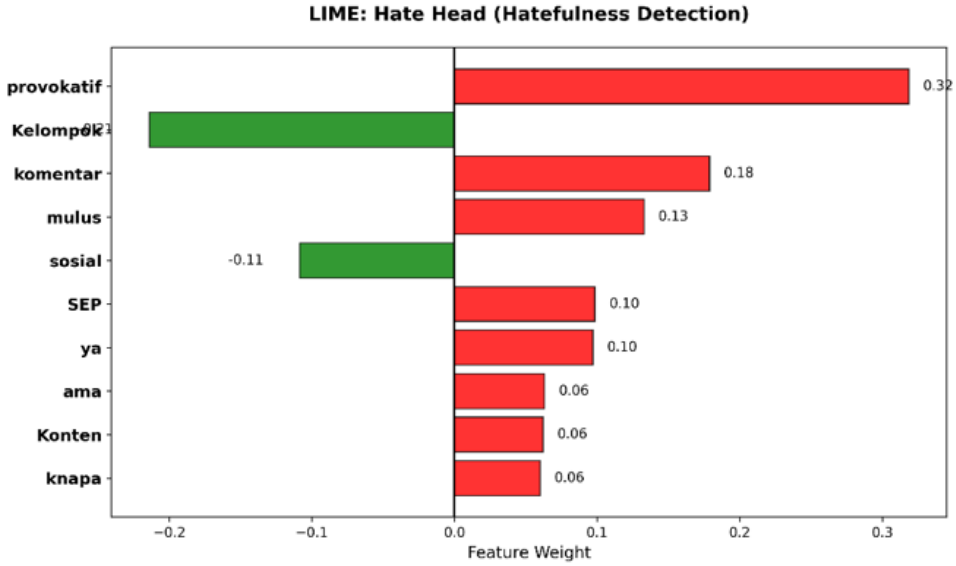

Figure 1. LIME Explanation for the Hatefulness Head

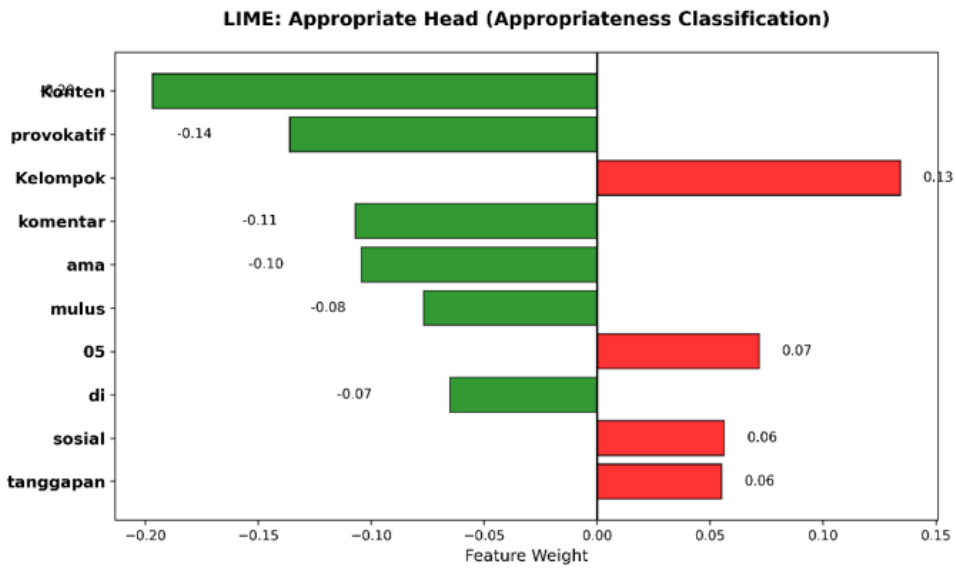

Figure 2. LIME Explanation for the Appropriateness Head on the Same Hateful Meme Instance

As observed in Figure 1, the "Hate" head assigns the highest positive weight to the word "provokatif" (0.32), followed by "komentar" (0.18) and "mulus" (0.13). Interestingly, the word "Kelompok" receives a strong negative weight (-0.21) in this head. This indicates that the hate head has specialized in detecting specific linguistic markers of social agitation, while actively disregarding neutral group identifiers. Conversely, when the exact same text is processed by the "Appropriate" head (Figure 2), the feature attribution distribution shifts dramatically. The word "Kelompok" flips to become the highest positive contributor (0.13), while "provokatif" receives a strong negative weight (-0.14). This stark contrast demonstrates that the appropriate head is highly sensi-

tive to general contextual descriptors and norm violations, whereas the hate head aggressively isolates specific provocative terms.

To validate the robustness of the model and ensure it does not produce false positives by arbitrarily flagging neutral text, we conducted a comparative analysis using a benign meme instance correctly classified as "Not Hate". Figure 3 illustrates the LIME explanation for the Hatefulness Head on this baseline instance.

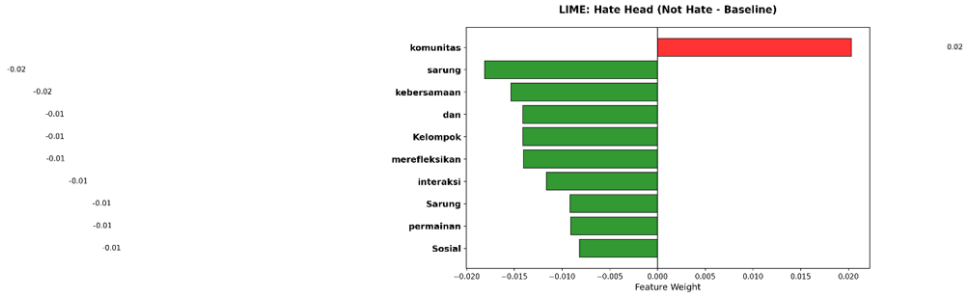

Figure 3. LIME Explanation for the Hatefulness Head on a Non-Hateful (Benign) Meme Instance

As demonstrated in Figure 3, the model assigns negligible weights (ranging strictly between -0.02 and 0.02) to all features, such as "komunitas" (0.02) and "kebersamaan" (-0.02). The absence of high-weight agitative features in this benign instance provides critical empirical evidence. It confirms that the model's detection of hate speech is strictly grounded in the presence of specific semantic cues rather than a biased predisposition to classify all enriched social media text as hateful.

Ultimately, this divergence in feature attribution across the three figures proves that the dual-head architecture does not merely duplicate the classification process. Instead, the shared IndoBERTweet backbone successfully learns to disentangle nuanced linguistic cues required for two distinct sociolinguistic tasks. The hate head specializes in detecting targeted social agitation, while the appropriate head focuses on broader contextual norms, all while maintaining a strict and reliable baseline against benign discourse.

3.4. Practical Implications and Limitations

The interpretability analysis validates the effectiveness of the dual-head architecture and demonstrates the immense value of XAI in auditing complex multitask models. For platform moderators and policymakers, understanding that the model relies on context-specific words like "provokatif" rather than mere profanity ensures that the system is capturing contextual malice rather than triggering false positives on benign slang or colloquialisms. This transparency is a crucial step toward deploying trustworthy AI for content moderation in low-resource language environments [20]. Furthermore, from a practical deployment perspective, the text-enriched strategy offers a significant computational advantage. By bypassing the need for heavy, dual-stream vision encoders (such as ViT or CLIP) and relying solely on enriched textual inputs, the proposed framework drastically reduces memory consumption and inference latency, making it highly suitable for real-time, large-scale moderation on resource-constrained platforms.

However, this approach is not without limitations. The primary constraint lies in its absolute dependency on the accuracy of the upstream text-enrichment models (Qwen2-VL-2B for OCR and Gemini 2.5 Flash for captioning). If a meme relies purely on visual irony with no extractable text, or if the captioning model hallucinates and generates an inaccurate description, the downstream IndoBERTweet classifier will inevitably fail to capture the true context. Additionally, the current study is strictly limited to the Indonesian language context and the specific cultural nuances of Indonesian social media. Future research should investigate the integration of lightweight vision encoders to handle purely visual memes and explore cross-lingual transferability to evaluate the model's robustness across different Southeast Asian languages.

3.5. Discussion

The primary contribution of this study lies in demonstrating that a text-enriched dual-head multitask architecture can achieve state-of-the-art performance in Indonesian hateful meme detection without the computational burden of full multimodal fusion. By elevating the Macro-F1 scores to 0.8120 and 0.8197 for hatefulness and appropriateness, respectively, this research validates the hypothesis that structured textual descriptors (OCR combined with AI-generated captioning) can effectively encapsulate the visual semantics necessary for implicit hate speech detection. This aligns with recent paradigms advocating for efficient representation learning in low-

resource languages, where direct pixel-level processing often yields diminishing returns relative to computational cost [12, 17].

The performance gains over the established baseline [1] are not merely artifacts of architectural adjustments, but rather the direct result of rigorous training optimizations tailored to the dataset's characteristics. The implementation of class-weighted loss and dynamic threshold calibration (shifting the decision boundary to 0.38 for appropriateness) directly mitigated the inherent class imbalance in the INDOMEME dataset. Furthermore, the application of label smoothing prevented the IndoBERTweet backbone from overfitting to highly polarized or sarcastic linguistic patterns. Instead, it forced the model to learn more robust, generalized semantic representations [18]. This regularization is particularly crucial in the Indonesian context, where code-mixing and cultural nuances frequently obscure explicit hate markers, making the model prone to false confidence without such constraints.

The most significant theoretical insight of this work emerges from the LIME-based interpretability analysis, which decodes the internal mechanics of the shared multitask backbone. Contrary to the assumption that dual-head models merely duplicate classification pathways, our empirical evidence reveals a sophisticated feature disentanglement. As demonstrated in the LIME visualizations, the hate-specific head aggressively isolates lexicons associated with social agitation (e.g., "provokatif"), while the appropriateness head prioritizes broader contextual descriptors and norm violations (e.g., "Kelompok"). This divergence substantiates that hard-parameter sharing in multitask learning enables the model to construct distinct, task-specific semantic subspaces from a unified representation. Consequently, this not only enhances predictive accuracy but also provides the decision transparency required for trustworthy AI auditing in sensitive content moderation scenarios [5, 16].

From a systems engineering perspective, the text-enrichment strategy presents a highly viable and novel alternative to computationally expensive vision-language models (e.g., CLIP or ViT-based architectures). By delegating visual understanding to specialized, pre-trained vision-language models (Qwen2-VL-2B and Gemini 2.5 Flash) and restricting the downstream classifier to a text-only transformer, the proposed framework drastically reduces inference latency and GPU memory footprint. This architectural decoupling makes the system highly scalable for real-time, large-scale content moderation on resource-constrained social media platforms, addressing a critical gap in practical AI deployment.

Despite these advancements, the framework exhibits inherent limitations tied to its sequential pipeline design. The model's performance is strictly bounded by the fidelity of the upstream text-enrichment phase; memes relying purely on visual irony, or those suffering from OCR hallucinations and inaccurate captioning, will inevitably lead to downstream classification failures. Furthermore, the current scope is confined to the sociolinguistic nuances of the Indonesian language. Future research should explore the integration of lightweight, parameter-efficient vision encoders (e.g., via Low-Rank Adaptation or adapters) to handle purely visual memes, alongside cross-lingual transfer learning to evaluate the architecture's robustness across broader Southeast Asian linguistic landscapes.

4. CONCLUSION

This study successfully implemented and optimized a text-enriched dual-head multitask learning approach for detecting hateful and inappropriate memes in Indonesian social media. By employing advanced training strategies, including cosine annealing, class-weighted loss, and dynamic threshold calibration, the optimized IndoBERTweet model achieved a Macro-F1 score of 0.8120 for hatefulness and 0.8197 for appropriateness, outperforming the established baseline. The integration of text-enrichment effectively captured complex semantic cues without the computational overhead of processing raw image pixels, resulting in a highly efficient and accurate moderation framework.

More importantly, the LIME-based interpretability analysis revealed that the dual-head architecture implicitly learns distinct feature representations for each task. Despite sharing the same textual backbone, the hate-specific head predominantly focuses on lexicons carrying social agitation, whereas the appropriateness head prioritizes general norm violations. This empirical finding underscores the value of multitask learning in building not only accurate but also transparent and trustworthy content moderation systems, providing platform moderators with clear, context-specific explanations for automated decisions.

However, this approach has limitations, primarily its absolute dependency on the accuracy of the upstream text-enrichment models for OCR and captioning. Memes that rely purely on visual irony without extractable text may not be accurately classified. Future research should investigate the integration of lightweight vision encoders

to handle purely visual memes and explore cross-lingual transferability to evaluate the model's robustness across different Southeast Asian languages.

DECLARATIONS

Author Contributions:
Selamet Riadi: Conceptualization, methodology, software, and writing – original draft preparation.
Emi Suryadi: Data curation and formal analysis.
Muhamad Masjun Efendi: Investigation and resources.
Bahtiar Imran: Validation and visualization.
Muhammad Zamroni Uska: Supervision, and writing – review and editing.

Funding:
This work was supported by Universitas Teknologi Mataram.

Conflicts of Interest:
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability:
The data that support the findings of this study are available from the corresponding author upon reasonable request.

AI Use Statement:
The authors utilized Qwen AI as a language editing and proofreading tool to enhance the linguistic quality and readability of the manuscript. The authors take full responsibility for the final content, accuracy, and scientific integrity of the paper. All AI-assisted outputs were thoroughly reviewed, verified, and approved by the authors.

Additional Information:
No additional information is available for this paper.

REFERENCES

[1] E. W. Pamungkas, C. S. Wahyuni, I. Amal, D. Purworini, and B. S. Rintyarna, "Decoding hate in memes: multimodal and multitask approaches for low-resource Indonesian social media," *PeerJ Comput. Sci.*, vol. 12, p. e3736, 2026.

[2] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," *Heliyon*, vol. 9, no. 8, 2023.

[3] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10660–10668.

[4] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *J. Big Data*, vol. 11, no. 1, p. 25, 2024.

[5] J. Zhang, K. Yan, and Y. Mo, "Multi-task learning for sentiment analysis with hard-sharing and task recognition mechanisms," *Information*, vol. 12, no. 5, p. 207, 2021.

[6] M. S. Hee, W.-H. Chong, and R. K.-W. Lee, "Decoding the underlying meaning of multimodal hateful memes," *arXiv Prepr. arXiv:2305.17678*, 2023.

[7] S. Ali et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Inf. Fusion*, vol. 99, p. 101805, 2023.

[8] A. Nirmal, A. Bhattacharjee, P. Sheth, and H. Liu, "Towards interpretable hate speech detection using large language model-extracted rationales," in *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, 2024, pp. 223–233.

[9] H. Kibriya, A. Siddiqa, W. Z. Khan, and M. K. Khan, "Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification," *Comput. Electr. Eng.*, vol. 116, p. 109153, 2024.

[10] D. Mittal and H. Singh, "Enhancing hate speech detection through explainable ai," in *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, IEEE, 2023, pp. 118–123.

[11] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, "Mapping memes to words for multimodal hateful meme classification," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, 2023, pp. 2824–2828.

[12] E. Fetahi, A. Susuri, M. Hamiti, Z. Kastrati, E. Canhasi, and A. Misini, "Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 82, 2025.

[13] L. Cao, X. Liu, and H. Shen, "Adaptable focal loss for imbalanced text classification," in *International Conference on Parallel and Distributed Computing: Applications and Technologies*, Springer, 2021, pp. 466–475.

[14] O. V. Johnson, C. Xinying, K. W. Khaw, and M. H. Lee, "ps-CALR: periodic-shift cosine annealing learning rate for deep neural networks," *IEEE Access*, vol. 11, pp. 139171–139186, 2023.

[15] T. Dehdarirad, "Evaluating explainability in language classification models: A unified framework incorporating feature attribution methods and key factors affecting faithfulness," *Data Inf. Manag.*, p. 100101, 2025.

[16] P. Kapil and A. Ekbal, "A transformer based multi task learning approach to multimodal hate speech detection," *Nat. Lang. Process. J.*, vol. 11, p. 100133, 2025.

[17] E. Hwang and V. Shwartz, "Memecap: A dataset for captioning and interpreting memes," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 1433–1445.

[18] H. Ren, Y. Zhao, Y. Zhang, and W. Sun, "Learning label smoothing for text classification," *PeerJ Comput. Sci.*, vol. 10, p. e2005, 2024.

[19] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[20] A. Rawat, S. Kumar, and S. S. Samant, "Hate speech detection in social media: Techniques, recent trends, and future challenges," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 16, no. 2, p. e1648, 2024.