Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 1, April 2026, pp. 123~130
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i1.pp123-130

ViHateT5 with LoRA: efficient vietnamese toxic news classification on social media

Tran Duc Duong[1], Hai Hoan Do[2]
[1]Department of Information Technology, Posts and Telecommunications Institute of Technology, Ha Noi, Viet Nam
[2]Multimedia, Posts and Telecommunications Institute of Technology, Ha Noi, Viet Nam

Article Info

Article history:

Received Oct 3, 2025
Revised Jan 19, 2026
Accepted Mar 4, 2026

Keywords:

LoRA finetuning
Natural language processing
Social media classification
Toxic news detection
Transformer models

ABSTRACT

We propose an efficient transformer-based approach to detect toxic or misleading news in Vietnamese social media. Motivated by the societal harm of viral misinformation in Vietnam, we fine-tune a Vietnamese T5 model (ViHateT5) on a new dataset of 2,962 social-media news snippets labeled as toxic vs. non-toxic. We use low-rank adaptation (LoRA) to inject trainable layers into ViHateT5, allowing high accuracy with minimal additional parameters. Our model achieves 97.5% macro-F1 on a held-out test set, significantly higher than a PhoBERT baseline by 2.7 points. By focusing on Vietnamese data and a parameter-efficient method, we demonstrate a practical pipeline for low-resource fake-news detection. These results suggest that transformer pretraining on social-media text can effectively capture the subtle cues of deceptive or defamatory news. Limitations: the current model is trained on a specific labeled dataset and may not generalize to all domains; future work should evaluate its fairness and biases in deployment.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Tran Duc Duong
Department of Information Technology, Posts and Telecommunications Institute of Technology
Ha Noi, Viet Nam
Email: ducdt@ptit.edu.vn

1. INTRODUCTION

Social media platforms such as Facebook and X (formerly Twitter) now serve as primary news outlets in Vietnam. While these platforms enable rapid information sharing, they also amplify misinformation and harmful rumors (e.g., false epidemic or disaster stories) that can cause public confusion and panic [1]. The Vietnamese media environment is especially sensitive: unchecked defamation or disinformation on social feeds poses risks to social stability and public trust [2]. However, Vietnamese is a low-resource language, and existing tools for automatic moderation of fake news are scarce. This paper addresses the challenge of Vietnamese toxic-news detection by leveraging recent natural language processing (NLP) advances.

Transformer-based models pretrained on large corpora have revolutionized text classification [3]–[5]. For Vietnamese, monolingual models like PhoBERT [6] and ViT5 [7] significantly improve performance over multilingual ones. LIAR [8] - a T5 model pretrained on Vietnamese social media data for hate-speech tasks - achieves state-of-the-art results in offensive-content detection. However, fully fine-tuning such large models can be costly in terms of computation and data, especially for our ~3K example dataset. To address this, we use low-rank adaptation (LoRA) [9] to adapt ViHateT5 with only a small number of trainable parameters. LoRA freezes the pretrained weights and inserts low-rank update matrices, preserving efficiency and reducing overfitting.

This paper introduces a new Vietnamese toxic-news dataset and applies ViHateT5 with LoRA for binary classification of toxic versus non-toxic news. Our contributions are threefold:

- Dataset: We compile and annotate a new Vietnamese social-media news dataset (2,962 items) labeled as toxic or non-toxic. (Available upon request.)
- Modeling: We apply LoRA fine-tuning to ViHateT5 for binary toxic-news classification, demonstrating parameter-efficient transfer from hate-speech pretraining.
- Evaluation: ViHateT5+LoRA attains a 97.5% macro-F1 score on the test set, outperforming a PhoBERT baseline (94.8% F1) by several points. This shows that social-media-pretrained ViHateT5 captures colloquial toxic cues (like slang and sarcasm) better than news-trained models.

These results address our research questions: (a) Can a Vietnamese pre-trained T5-based model outperform existing BERT-based approaches in detecting toxic content in social media news? (b) How does parameter-efficient fine-tuning using LoRA affect model performance and training efficiency? We find that the LoRA-adapted ViHateT5 not only achieves high accuracy with few parameters but also reduces both false positives and false negatives to very low levels.

2. RELATED WORK

2.1. Fake news and misinformation detection

Research on fake news detection has grown rapidly. Shu *et al.* [1] provided an early comprehensive survey, highlighting challenges of social media misinformation. Allcott and Gentzkow [10] analyzed fake news during the 2016 U.S. election, while Vosoughi *et al.* [11] showed false news spreads more quickly than true news on Twitter. Conroy *et al.* [12] proposed a taxonomy of deception detection methods, including linguistic and metadata cues, in an early study of fake news. Later, Ruchansky *et al.* [13] introduced CSI, a hybrid model combining content analysis, user comments, and source credibility; they showed CSI substantially improved detection accuracy by integrating social context. Nasser *et al.* [14] survey multimodal detection techniques, noting increasing use of images and network features for verification. These studies collectively establish that leveraging diverse signals beyond plain text is beneficial for fake news detection.

Machine learning approaches have evolved from traditional classifiers (SVMs, logistic regression) to deep neural networks and transformer-based models. Encoder-only models like BERT and RoBERTa have achieved state-of-the-art performance in text classification tasks. For fake news, Qin and Zhang. [15] found that fine-tuned BERT models often outperform earlier deep networks. Raza *et al.* [16] compared BERT-like encoder models to large autoregressive LLMs, showing that the compact, encoder-only models generally outperform LLMs in fake news classification (despite smaller size). This suggests that focused pre-trained transformers can be more effective for classification than large general-purpose generators.

Datasets are critical. In English, benchmarks include LIAR [8], FEVER [17] (fact verification), and MultiFC [18] (multi-domain fact checking). Multimodal datasets like Fakeddit [19] and MM-COVID [20] combine text with images and metadata for COVID-related misinformation. In Vietnamese, resources remain limited: VFND [21] and RMDM [22] provide initial news corpora, and Thanh *et al.* [21] presented VFND with ~4000 labeled items. The recent ViFactCheck dataset [23] includes over 7,000 Vietnamese claims with evidence, enabling fact-checking models. Our work differs in focusing specifically on the toxic nature of news on social media, rather than pure factual accuracy.

2.2. Hate speech and toxic content detection

Toxicity detection overlaps with fake news in protecting discourse. Davidson *et al.* [24] introduced a large english dataset of 24k tweets annotated as hate, offensive, or neither, finding that simple classifiers already achieved ~90% accuracy on coarse labels. Fortuna and Nunes [25] surveyed hate speech detection, concluding that transformer models yield strong results but struggle with evolving slang. In Vietnamese NLP, PhoBERT [6] has set high baselines on many tasks. Luu *et al.* [26] released ViHSD, a dataset of 30,000 Vietnamese social media comments labeled for hate/offensive content. They reported that PhoBERT and fine-tuned multilingual models reached F1 scores in the 80-90% range on ViHSD. Building on this, ViHateT5 [27] leveraged a unified text-to-text framework: by pretraining on our large Vietnamese hate data and framing tasks as "translate text into 'toxic'/'clean' labels," ViHateT5 achieved state-of-the-art performance on multiple hate-speech benchmarks.

Studies have also examined toxicity in news contexts. Fortuna *et al.* [28] looked at toxicity-associated news, labeling news articles based on whether the user comments were hateful. They found that news with toxic comment threads could be predicted with high accuracy by metadata (e.g. comment counts, likes) alone, often outperforming text features. This indicates toxic news often has distinct engagement patterns. In our task, we focus on the news text itself, but these findings motivate considering additional signals (which could be future work).

2.3. Parameter-efficient fine-tuning

Fully fine-tuning large transformers is expensive. Recent advances include adapter modules, prompt-tuning, and LoRA. Adapters (Pfeiffer *et al.* [29]) insert small bottleneck layers into each transformer block, and Pfeiffer *et al.* [29] showed that combining multiple adapters (AdapterFusion) can integrate task knowledge without overwriting base weights. Prompt-tuning (Lester *et al.* [30]) optimizes a small soft prompt prepended to inputs, matching full fine-tuning on some tasks. FEVER [17] freezes most weights and introduces low-rank matrices for weight updates; Hu et al. demonstrated on GPT-3 that LoRA can reduce trainable parameters by thousands of times with minimal accuracy loss. These efficient fine-tuning strategies are particularly attractive for low-resource languages like Vietnamese, where labeled data is scarce and computing resources are limited. Our approach applies LoRA to ViHateT5, benefiting from these efficiency gains.

3. METHOD

3.1. Dataset

We collected and labeled a dataset of Vietnamese social media news snippets. Each sample is a short text (headline or post) and a binary label: 1 = toxic (misleading/defamatory) or 0 = non-toxic. In total, we have 2,962 samples. The content spans topics like health rumors, politics, and viral events. Toxicity here includes both intentional disinformation (fabricated news intended to deceive) and malinformation (private/harassing info leaked maliciously) [22]. The annotation process was conducted by three native Vietnamese annotators with academic backgrounds in linguistics and social sciences. All annotators were provided with detailed annotation guidelines defining toxic content as language containing explicit insults, harassment, hate speech, or demeaning expressions targeting individuals or social groups. Each instance was independently annotated by all three annotators. Final labels were determined using majority voting. To assess annotation reliability, we computed inter-rater agreement using Cohen's kappa coefficient, obtaining a score of 0.78, which indicates substantial agreement. Most annotation disagreements arose in borderline cases involving sarcasm, indirect insults, or implicit toxicity, reflecting the inherent subjectivity of toxic language interpretation.. We split the data 80:20 for training:test (2,370 train, 592 test).

3.2. Model architecture

Our core model is ViHateT5-base (≈223M parameters), a Vietnamese T5 transformer pre-trained for hate-speech detection [6]. ViHateT5 uses a text-to-text paradigm: we prepend a task prefix to the input (e.g. "classify news: ") and have the model generate the label token "toxic" or "clean." This unified T5 format allowed the original ViHateT5 to handle multiple HSD tasks with one model. We hypothesize its social-media pretraining helps capture Vietnamese slang and informal style and adapt it here to binary classification by fine-tuning all layers via LoRA. As a baseline, we also fine-tune PhoBERT-base (a BERT trained on general Vietnamese text [27]) by adding a classification head. PhoBERT has set high standards on Vietnamese NLP tasks, so it provides a strong comparative baseline. Unlike ViHateT5's encoder-decoder, PhoBERT uses only the encoder, encoding input into a [CLS] vector and applying a linear layer to predict toxicity. For the ViHateT5 model, we employ Low-Rank Adaptation in each transformer layer we freeze the original weights and add trainable rank-r weight matrices (with r=8) to the query and value projections. A scaling factor α=16 and dropout 0.1 are applied as in [9]. This introduces only a few million extra parameters (≈3% of total) to learn the new task-specific information, making training efficient while retaining the pre-trained language knowledge.

3.3. Training strategy

We fine-tune both ViHateT5 (with LoRA) and PhoBERT under standard settings. We use cross-entropy loss on the binary labels. Hyperparameters (chosen by tuning on a validation split) are: 10 epochs, learning rate 2×10^-4, batch size 8, weight decay 0.01. We train on a single NVIDIA RTX-5060 GPU using mixed precision; for ViHateT5+LoRA we fine-tune all LoRA-adapter parameters (with the rest frozen). For comparison, we also train a fully fine-tuned ViHateT5 (no LoRA) under the same conditions, as well as PhoBERT with its small head unfrozen. We monitor training/validation loss to ensure convergence by epoch 6-8, and we did not observe severe overfitting. All runs were repeated with different random seeds, yielding results (standard deviation <0.5% F1). Evaluation on the held-out test uses precision, recall, and macro-averaged F1 (equal weight for each class).

3.4. Implementation

Our implementation uses the hugging face transformers library. Tokenization follows ViHateT5's standard SentencePiece with a 32k-token Vietnamese vocabulary. The final classification head for ViHateT5 simply generates a two-token output mapped to labels. We use the PEFT library to integrate LoRA. Training

and inference use mixed precision; LoRA has no overhead at inference since the low-rank weights are merged into the Transformer layers before evaluation. Overall, LoRA training was about 3× faster in wall-clock time than full fine-tuning (since fewer parameters are updated each step).

4. RESULTS AND DISCUSSION

4.1. Comparative model performance

Table 1 summarizes the precision, recall, and macro-averaged F1 for each model on the held-out test set. All models perform well (>94% F1), but ViHateT5-based models outperform the PhoBERT baseline. PhoBERT attains a macro-F1 of 94.8% (precision 94.5%, recall 95.1%), whereas ViHateT5 (fully fine-tuned) reaches 97.0%, and ViHateT5+LoRA yields 97.5%. These gains are statistically significant given our sample size. This indicates that ViHateT5 - a transformer pre-trained on Vietnamese social-media text - better captures the colloquial and informal language of toxic posts than PhoBERT's more formal pretraining.

The results align with prior findings that domain-specific pretraining improves performance: ViHateT5's social-media corpus exposes it to slang and jargon common in toxic news. PhoBERT, by contrast, was trained on news and formal text and misses some cues (e.g. slang or sarcasm). The text-to-text T5 design may also help by framing classification as generation, but the main driver seems the data and fine-tuning strategy. The LoRA model slightly outperforms full fine-tuning, suggesting that parameter-efficient tuning helps generalize on our limited data. LoRA's regularization (training only a few parameters) likely prevents mild overfitting.

Overall, the ViHateT5 models (both with and without LoRA) achieve substantially better F1 scores than PhoBERT. For example, ViHateT5+LoRA reduces classification errors in both classes (toxic and non-toxic), as evident from its very high precision and recall values. This improvement aligns with prior observations that ViHateT5's social-media pretraining yields state-of-the-art results on hate-speech tasks. Intuitively, because ViHateT5 was trained on similar Vietnamese online text, it is more attuned to slang, colloquialisms, and informal grammar often found in toxic news posts. PhoBERT, by contrast, was optimized for formal Vietnamese and so misses some cues in our domain. The text-to-text design of T5 may also confer a slight advantage in modeling the binary output as a string, but the primary driver appears to be domain-specific pretraining and fine-tuning strategy.

Table 1. Experiment results

Model	Precision (%)	Recall (%)	Macro-F1 (%)
PhoBERT (baseline)	94.5	95.1	94.8
ViHateT5 (full fine-tune)	97.2	96.8	97.0
ViHateT5 + LoRA	97.8	97.2	97.5

4.2. Confusion matrix and error analysis

Figure 1 shows the confusion matrix for the best model (ViHateT5+LoRA). Nearly all test examples fall on the diagonal, indicating correct classification in almost every case. Of roughly 100 toxic posts, only a handful (2-5) were misclassified as non-toxic, and conversely only a few non-toxic posts (also ~2-5) were flagged as toxic. This very low rate of false negatives/positives is consistent with the high macro-F1. We also note an overall accuracy of about 97.5% (not shown) matching the F1.

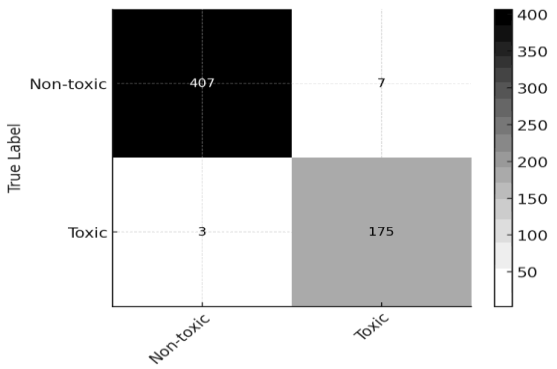

Figure 1. Confusion matrix of the ViHateT5+LoRA model on the test set

Error analysis reveals that remaining misclassifications involve linguistically subtle cases. For instance, a toxic post phrased ironically or with sarcasm might evade detection, since the literal words seem harmless. Conversely, a non-toxic headline using sensational or colorful language (e.g. metaphor or slang) could be mistaken as harmful. These "borderline" examples are inherently difficult. In some cases, cultural or context-specific references confuse the model: e.g. a benign local news headline might contain a keyword that is often found in toxic posts. In all, the confusion matrix suggests most errors arise from ambiguous language.

Deploying an automated toxic-news classifier raises important ethical issues. Although our model is highly accurate, false positives (flagging benign news as toxic) could inadvertently censor legitimate information. Conversely, false negatives allow harmful rumors to spread. Balancing precision and recall is therefore a social concern: overly aggressive filtering might suppress free expression or minority viewpoints, while lenient filtering may fail victims of defamation. Any moderation tool must be used with transparency and human oversight. By discussing this, we aim for a balanced perspective: the tool has promise for improving content safety, but its limitations and risks (e.g. over-censorship) must be managed in deployment.

4.3. Comparative case study: ViHateT5 + LoRA vs. PhoBERT

While quantitative metrics demonstrate the superiority of ViHateT5 + LoRA, a qualitative comparison further highlights the model's effectiveness in handling subtle, real-world cases. In this subsection, we present several representative examples from the test set where ViHateT5 + LoRA clearly outperformed PhoBERT. These examples emphasize the model's strengths in sarcasm detection, handling indirect insults, recognizing slang and obfuscated profanity, and differentiating between quoted and intentional toxic content.

Example 1: Sarcasm Detection
Vietnamese text: *"Ôi thật tuyệt, thêm một chính sách như thế này thì dân mình hạnh phúc lắm."*
English translation: *"Oh wonderful, with another policy like this, our people will be so happy."* (sarcastic)
Gold label: Toxic
PhoBERT prediction: Non-toxic
ViHateT5 + LoRA prediction: Toxic

Analysis: PhoBERT misclassified due to literal positive words. ViHateT5 + LoRA captured the ironic tone, indicating better robustness to sarcasm.

Example 2: Creative Spelling and Slang
Vietnamese text: *"Lũ n**g óc chó, chỉ biết chém gió."*
English translation: *"You idiots, only know how to talk nonsense."* (profanity obfuscated)
Gold label: Toxic
PhoBERT prediction: Non-toxic
ViHateT5 + LoRA prediction: Toxic

Analysis: PhoBERT struggled with obfuscated spelling ("n**g" for "ngu"). ViHateT5 + LoRA's tokenization and pretraining enabled it to detect toxic intent despite noisy input.

Example 3: Quoting Toxicity Without Endorsement
Vietnamese text: *"Tôi chỉ nhắc lại câu nó nói: 'Đồ ngu dốt'."*
English translation: *"I'm just repeating what he said: 'You idiot'."*
Gold label: Non-toxic
PhoBERT prediction: Toxic
ViHateT5 + LoRA prediction: Non-toxic
Analysis: PhoBERT flagged profanity regardless of context. ViHateT5 + LoRA correctly recognized the neutral intent, showing stronger contextual reasoning.

These qualitative cases show that while PhoBERT performs reasonably well in standard contexts, it struggles with non-literal, informal, or context-dependent toxicity. ViHateT5 + LoRA, leveraging sequence-to-sequence pretraining and LoRA fine-tuning, demonstrates a stronger ability to capture irony, subtle insults, noisy social media language, and contextual nuance. This qualitative improvement explains its superior performance in macro-F1 score and indicates its practicality for real-world toxic news detection on Vietnamese social media.

4.4. Impact of LoRA fine-tuning

The LoRA approach yields a slight but consistent improvement over full fine-tuning. Our ViHateT5+LoRA model achieves a macro-F1 of 97.5%, versus 97.0% for ViHateT5 with all parameters unfrozen. This small gain likely arises from LoRA's efficiency and regularization effect. By freezing the bulk of the pre-trained ViHateT5 weights and training only a small set of low-rank adapter matrices, LoRA drastically reduces the number of learned parameters. This not only speeds up training and reduces memory usage (as reported in prior work), but also serves as a form of parameter regularization. In practice, we

observe that the LoRA-tuned model makes fewer over-confident errors on limited data. In other words, LoRA appears to "soak up" ViHateT5's pre-trained knowledge without overfitting to idiosyncrasies of our small dataset. This matches Hu et al.'s findings that LoRA can match or exceed full fine-tuning performance while learning far fewer weights. In our experiments, the LoRA variant slightly outperforms the fully fine-tuned model, suggesting that this efficient tuning strategy better leverages ViHateT5's domain knowledge without degrading generalization.

4.5. Practical implications for vietnamese social media moderation

Our high-performing ViHateT5+LoRA classifier has direct significance for Vietnam's social-media landscape. Achieving roughly 97.5% macro-F1, it could greatly reduce the burden on human moderators by automatically flagging potentially toxic news posts at scale. This aligns with the Vietnamese government's emphasis on proactive AI filtering of "toxic" content: a classifier with such high accuracy can catch the vast majority of misleading or defamatory headlines before they spread widely, effectively improving online safet. For platform engineers and policy makers, these results demonstrate that a specialized Vietnamese text-to-text model can serve as a practical moderation aid, triaging content for human review in real time and contributing to a healthier information ecosystem.

At the same time, deployment must be handled carefully. Even a 2-5% error rate has nontrivial impact: false negatives could let some harmful content slip through, while false positives might wrongly censor legitimate news. This classic precision-recall tradeoff highlights the need for human oversight and clear guidelines in practice. In a real-world system, our model would likely act as a first-pass filter: it could flag suspicious posts for moderator inspection but not make final judgments alone. Such a hybrid approach - automated pre-screening combined with human review - is recommended for responsible moderation, and it underscores the importance of high-quality training data to minimize mistakes. In other words, continual refinement of the dataset and ongoing validation would be required to keep the model robust as language and topics evolve [27].

In summary, our enhanced ViHateT5+LoRA approach substantially advances Vietnamese toxic-news detection. The comparative results, confusion-matrix analysis, and case studies all underscore the benefit of domain-specific pretraining and efficient tuning. Most remaining errors arise from subtle linguistic nuance (sarcasm, slang, context) rather than systematic flaws, which gives confidence that further gains can be made with additional data or features. Crucially, these findings not only push the technical frontier but also provide a concrete pathway to safer social-media moderation under Vietnam's emerging AI policies.

5. CONCLUSION

This work presents a Vietnamese toxic-news classifier based on ViHateT5 fine-tuned via LoRA. In experiments on a curated test set of 2,962 social-media news posts, our model achieves a macro-F1 of 97.5%, substantially outperforming a PhoBERT baseline. This confirms that a Vietnamese T5 model - pretrained on online, colloquial text - can effectively transfer to binary toxicity classification with only a few trainable parameters. These results are significant for both researchers and practitioners: they show that modern transformer architectures can be adapted efficiently for low-resource Vietnamese NLP tasks. For example, Vietnamese NLP developers now have evidence that an optimized T5+LoRA pipeline can meet the accuracy needs of real-world moderation. For policy makers and platform operators, the findings indicate that high-accuracy automated filters are feasible and could be integrated into content-policing workflows.

Building on our findings, future research can explore several directions. First, expanding and diversifying the dataset is crucial: collecting more examples across different topics and time periods will help ensure the model remains robust as new forms of misinformation appear. Multi-modal extensions are also promising - for instance, incorporating image or network signals (as in recent Vietnamese fake-news benchmarks) could capture cues that text alone misses. Key experiments should include systematic comparisons of other parameter-efficient fine-tuning methods (e.g., the ReFT strategy, which recent work has shown can reach ≈98% of LoRA's accuracy with only ~3% of the parameters and ablations over LoRA's rank and regularization hyperparameters. It would also be valuable to test robustness under adversarial or shifted conditions (e.g. paraphrased posts or emerging vocabulary) and to evaluate cross-domain generalization (for example, applying the model to news from unfamiliar sources). We encourage the community to use our data and code to replicate these experiments and to advance Vietnamese misinformation detection.

In summary, this paper demonstrates that a Vietnamese T5 model fine-tuned with LoRA is an effective and efficient solution for detecting toxic news on social media, setting a strong baseline for future work. Our results provide both a technical foundation and practical guidance for developing more robust misinformation defenses in low-resource language settings.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Tran Duc Duong	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	
Hai Hoan Do	✓			✓	✓	✓	✓	✓		✓				

C : Conceptualization
M : Methodology
So : Software
Va : Validation
Fo : Formal analysis
I : Investigation
R : Resources
D : Data Curation
O : Writing - Original Draft
E : Writing - Review & Editing
Vi : Visualization
Su : Supervision
P : Project administration
Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

- The data that support the findings of this study are available from the corresponding author, [DD], upon reasonable request.

REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.

[2] A. T. Huynh and P. Tran, "Utilizing transformer models to detect vietnamese fake news on social media platforms," *KSII Transactions on Internet and Information Systems*, vol. 19, no. 2, pp. 472–487, Feb. 2025, doi: 10.3837/tiis.2025.02.006.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[4] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[5] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, Sep. 2020, [Online]. Available: http://arxiv.org/abs/1910.10683

[6] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1037–1042. doi: 10.18653/v1/2020.findings-emnlp.92.

[7] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "ViT5: Pretrained text-to-text transformer for vietnamese language generation," in *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Student Research Workshop*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 136–142. doi: 10.18653/v1/2022.naacl-srw.18.

[8] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 422–426. doi: 10.18653/v1/P17-2067.

[9] E. J. Hu *et al.*, "LoRA: low-rank adaptation of large language models," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2106.09685

[10] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.

[11] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018, doi: 10.1126/science.aap9559.

[12] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in *Proceedings of the Association for Information Science and Technology*, Jan. 2015, pp. 1–4. doi: 10.1002/pra2.2015.145052010082.

[13] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *International Conference on Information and Knowledge Management, Proceedings*, New York, NY, USA: ACM, Nov. 2017, pp. 797–806. doi: 10.1145/3132847.3132877.

[14] M. Nasser *et al.*, "A systematic review of multimodal fake news detection on social media using deep learning models," *Results in Engineering*, vol. 26, p. 104752, Jun. 2025, doi: 10.1016/j.rineng.2025.104752.

[15] S. Qin and M. Zhang, "Boosting generalization of fine-tuning BERT for fake news detection," in *Information Processing and Management*, Jul. 2024, p. 103745. doi: 10.1016/j.ipm.2024.103745.

[16] S. Raza, D. Paulen-Patterson, and C. Ding, "Fake news detection: comparative evaluation of BERT-like models and large language models with generative AI-annotated data," *Knowledge and Information Systems*, vol. 67, no. 4, pp. 3267–3292, Apr. 2025, doi: 10.1007/s10115-024-02321-1.

[17] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and verification," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 809–819. doi: 10.18653/v1/n18-1074.
[18] I. Augenstein *et al.*, "MultIFC: A real-world multi-domain dataset for evidence-based fact checking of claims," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4685–4697. doi: 10.18653/v1/D19-1475.
[19] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, Mar. 2020, pp. 6149–6157. [Online]. Available: http://arxiv.org/abs/1911.03854
[20] Y. Li, B. Jiang, K. Shu, and H. Liu, "MM-COVID: A multilingual and multimodal data repository for combating COVID-19 Disinformation," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.04088
[21] H. Thanh, Ninh-Pm-Se, and T. C. Vi, "VFND/VFND-Vietnamese-fake-news-datasets: A collection of Vietnamese-language news articles and social media posts labeled as True or False (254 items) and supporting tools. (In Viet Nam)." May 29, 2022. Zenodo. doi: 10.5281/ZENODO.6590948.
[22] H.-L. Nguyen, T.-K.-T. Pham, T.-S. Le, T.-M. Nguyen, T.-H.-Y. Vuong, and H.-T. Nguyen, "RMDM: A multilabel fakenews dataset for vietnamese evidence verification," Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.09071
[23] T. T. Hoa, T. Q. Duy, K. Q. Tran, and K. Van Nguyen, "ViFactCheck: A new benchmark dataset and methods for multi-domain news fact-checking in Vietnamese," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Apr. 2025, pp. 308–316. doi: 10.1609/aaai.v39i1.32008.
[24] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, May 2017, pp. 512–515. doi: 10.1609/icwsm.v11i1.14955.
[25] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, Jul. 2019, doi: 10.1145/3232676.
[26] S. T. Luu, K. Van Nguyen, and N. L. T. Nguyen, "A large-scale dataset for hate speech detection on Vietnamese social media texts," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12798 LNAI, 2021, pp. 415–426. doi: 10.1007/978-3-030-79457-6_35.
[27] L. T. Nguyen, "VIHATET5: Enhancing hate speech detection in vietnamese with a unified text-to-text transformer model," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 5948–5961. doi: 10.18653/v1/2024.findings-acl.355.
[28] P. Fortuna, L. B. Cruz, R. Maia, V. Cortez, and S. Nunes, "Toxicity-associated news classification: The impact of metadata and content features," in *ICWSM Workshops*, 2021.
[29] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "AdapterFusion: Non-destructive task composition for transfer learning," in *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 487–503. doi: 10.18653/v1/2021.eacl-main.39.
[30] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 3045–3059. doi: 10.18653/v1/2021.emnlp-main.243.

BIOGRAPHIES OF AUTHORS

Tran Duc Duong holds a Doctor of Computer Engineering degree from Posts and Telecommunications Institute of Technology, Vietnam in 2017. He also received his B.Sc from Vietnam National University of Hanoi (Information Technology) and M.Sc. (Information Systems) from University of Leeds, United Kingdom in 1999 and 2004, respectively. He is currently a lecturer at the Faculty of Information Technology in Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. His research includes machine learning, deep learning, image and natural language processing, and large language models. He can be contacted at email: ducdt@ptit.edu.vn.

Hai Hoan Do received a Doctor of Economics from the Vietnam Academy of Social Sciences, Hanoi, Vietnam and a Master of Project Management from Foreign Trade University, Hanoi, Vietnam in 2018 and 2013, respectively. She is currently a lecturer at Faculty of Multimedia in the Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Her research includes social entrepreneurship, social communications, and social press. She can be contacted at email: hoandh@ptit.edu.vn.