Detecting Sophisticated Fake Reviews on E-Commerce Platforms Using Adversarial Transformer Networks

Sabar Aritonang Rajagukguk
Management Department, Binus Online Learning, Bina Nusantara University, Jakarta, Indonesia
sabar.aritonang@binus.ac.id (corresponding author)

Dedy Sofyan
Aspirasi Hidup Indonesia Corporation, Jakarta, Indonesia
dedysofyanm@gmail.com

Received: 14 July 2025 | Revised: 12 September 2025 and 6 October 2025 and | Accepted: 9 October 2025

 DOI: https://doi.org/10.48084/etasr.13369

ABSTRACT

The proliferation of Artificial Intelligence (AI)-generated fake reviews poses an unprecedented threat to the integrity of e-commerce platforms, particularly in developing markets where regulatory frameworks remain nascent. This study proposes an adversarial transformer network framework specifically designed to detect sophisticated fake reviews on Indonesian e-commerce platforms. We developed a novel adversarial training architecture that pairs a Bidirectional Encoder Representations from Transformers (BERT)-based classifier model with a generator capable of producing human-like fake reviews, creating an iterative optimization process that enhances detection robustness. The scientific novelty of this work is threefold: (i) architectural innovation, through the integration of IndoBERT as a discriminator with a fine-tuned Generative Pre-trained Transformer (GPT)-based generator in a competitive adversarial loop; (ii) linguistic innovation, by embedding Indonesian-specific preprocessing (slang handling, code-mixed normalization, emoticon filtering) to address multilingual and culturally diverse contexts; and (iii) training innovation, by introducing gradient penalty mechanisms and iterative adversarial updates that enhance robustness against Large Language Model (LLM)-generated reviews. Together, these contributions distinguish our framework from prior adversarial Natural Language Processing (NLP) approaches that primarily focused on English-language data and lacked local linguistic customization. To the best of our knowledge, this represents the first adversarial transformer framework tailored for Indonesian e-commerce fake review detection. Using a comprehensive dataset of 50,000 authentic reviews collected from major Indonesian e-commerce platforms (Tokopedia and Shopee) and 25,000 AI-generated fake reviews, our methodology achieved significant improvements over traditional detection methods. The adversarial framework demonstrated superior performance with an accuracy of 94.3%, precision of 93.8%, recall of 94.7%, and F1-score of 94.2%, outperforming baseline BERT models by 8.7% in accuracy. Our approach addresses the critical challenge of detecting increasingly sophisticated AI-generated fake reviews while providing insights into the unique linguistic patterns of Indonesian online commerce discourse. The findings contribute to both the theoretical understanding of adversarial learning in NLP and practical applications for maintaining trust in digital marketplaces.

Keywords-adversarial networks; fake review detection; Bidirectional Encoder Representations from Transformers (BERT); e-commerce; Indonesian market; transformer models; Natural Language Processing (NLP); digital trust

I. INTRODUCTION

The digital transformation of commerce has fundamentally altered consumer behavior, with online reviews serving as critical decision-making factors for purchasing [1]. Indonesian e-commerce markets experienced 30% growth in 2023, where reviews have become integral to platform trust mechanisms [2]. However, the increasing sophistication of Artificial Intelligence (AI) tools enables the generation of fake reviews that are virtually indistinguishable from authentic consumer feedback [3].

Traditional fake review detection methods have relied on linguistic pattern analysis, behavioral signals, and statistical anomaly detection [4]. While effective against manually crafted fake reviews, these approaches struggle when confronting AI-generated content that closely mimics human writing patterns, achieving accuracy rates below 70% in such contexts [5]. Early

studies primarily focused on linguistic features and statistical anomalies, achieving only moderate success, whereas more recent transformer-based approaches, including Bidirectional Encoder Representations from Transformers (BERT) variants, capture contextual relationships and semantic nuances, achieving accuracy rates exceeding 87% [6].

The Indonesian e-commerce ecosystem presents unique challenges for fake review detection due to its multilingual environment, diverse cultural contexts, and rapidly evolving digital commerce practices [7]. Major platforms such as Tokopedia and Shopee host millions of reviews daily, creating fertile ground for automated fake review generation that can significantly impact consumer trust and market dynamics [8]. Despite this, most existing research focuses on English-language platforms, with limited attention to robust detection mechanisms tailored for AI-generated content in emerging markets with unique linguistic characteristics [9, 10].

Adversarial training has emerged as a promising approach to address these challenges by creating competitive learning environments where detection models continuously adapt to increasingly sophisticated fake content generation [11, 12]. Generative Adversarial Networks (GANs) applied to fake review detection show improved performance through adversarial learning mechanisms, although challenges remain in generating sufficient training data and ensuring consistent accuracy. The intersection of adversarial training and transformer-based models, particularly in non-English markets, remains underexplored, offering significant potential for both theoretical advancement and practical applications [13].

This study aims to enhance detection of sophisticated AI-generated fake reviews in Indonesian e-commerce platforms using adversarial transformer networks. The scientific novelty lies in three aspects: (i) architectural innovation through integrating IndoBERT as a discriminator and fine-tuned Generative Pre-trained Transformer-2 (GPT-2) generator into competitive adversarial loops; (ii) linguistic adaptation via Indonesian-specific preprocessing and IndoBERT tokenizer, enhancing detection in multilingual environments; and (iii) training innovation by introducing gradient penalty mechanisms and ablation validation, significantly improving robustness. This combination distinguishes our approach from prior adversarial Natural Language Processing (NLP) models, which were primarily developed for English-language datasets without local linguistic customization.

This research makes three clear contributions to the field of adversarial NLP. First, we propose an architectural integration of IndoBERT and GPT-based generator tailored for Indonesian review detection. Second, we introduce Indonesian-specific linguistic preprocessing that captures colloquialisms, mixed-language expressions, and cultural nuances often overlooked in global models. Third, we implement a training scheme with adversarial robustness mechanisms, including gradient penalties and ablation-based validation. These combined contributions explicitly differentiate our framework from prior adversarial text classification models, which largely focused on English corpora and general transformer baselines.

II. METHODOLOGY

A. Study Design and Dataset Construction

This study employs an experimental design utilizing adversarial training to develop robust fake review detection systems. The research framework consists of three primary components: authentic review data collection from Indonesian e-commerce platforms, AI-generated fake review synthesis, and adversarial neural network training.

A comprehensive dataset comprising 75,000 product reviews was constructed, collected from two major Indonesian e-commerce platforms: Tokopedia and Shopee. The authentic review dataset contains 50,000 verified reviews spanning diverse product categories including electronics, fashion, home goods, beauty products, and food items. Reviews were collected over a six-month period from January to June 2024, ensuring temporal diversity and reducing selection bias.

Authentic reviews were selected based on several verification criteria requiring verified purchase status, substantial textual content (minimum 20 characters), and associated user profile legitimacy indicators. Reviews flagged by platform moderation systems were excluded, and linguistic filters ensured Indonesian language content. The dataset encompasses reviews ranging from 20 to 500 characters in length, representing typical consumer feedback patterns.

B. Synthetic Fake Review Generation

To supplement the dataset with sophisticated fake reviews, GPT-3.5-turbo was employed to generate 25,000 synthetic fake reviews. The generation process utilized carefully crafted prompts designed to produce contextually appropriate, product-specific fake reviews mimicking authentic Indonesian consumer language patterns. Generation prompts incorporated product categories, sentiment variations, and stylistic diversity, ensuring realistic fake review characteristics.

Representative prompt templates included: "Write a short Indonesian-language review (30-50 words) for a smartphone purchased on Tokopedia. Include common slang, sentiment (positive/negative), and mixed Indonesian-English expressions." Prompts varied by domain (electronics, fashion, food) ensuring coverage of Indonesian cultural and linguistic diversity.

C. Data Preprocessing and Feature Engineering

Comprehensive preprocessing operations were applied, ensuring data quality and model compatibility. Text normalization procedures included lowercase conversion, punctuation standardization, and whitespace regularization. Indonesian-specific preprocessing addressed common abbreviations, colloquialisms, and mixed-language expressions prevalent in local e-commerce contexts.

Tokenization utilized IndoBERT tokenizer, specifically designed for Indonesian text processing. The maximum sequence length truncation at 512 tokens maintained computational efficiency while preserving essential review content. Special attention was given to handling emoticons, product-specific terminology, and Indonesian slang expressions common in e-commerce reviews.

D. Adversarial Network Architecture

The proposed adversarial framework consists of two primary neural networks: generator and discriminator. The generator employs a fine-tuned GPT-2 architecture adapted for Indonesian text generation, whereas the discriminator utilizes a BERT-base model specifically pre-trained on Indonesian text corpora. This adversarial configuration creates competitive learning environments where both networks continuously improve through iterative training.

Figure 1 presents the detailed architecture of the adversarial transformer network. The diagram illustrates the end-to-end pipeline: authentic and AI-generated reviews as inputs, Indonesian-specific preprocessing, generator (GPT-2) producing synthetic reviews, discriminator (IndoBERT) classifying fake versus authentic reviews, training configuration (AdamW optimizer, learning rate 2×10^{-5}, batch size 16, 100 epochs, gradient clipping), and output metrics (accuracy, precision, recall, F1-score, AUC). The feedback loop between generator and discriminator illustrates adversarial learning processes.

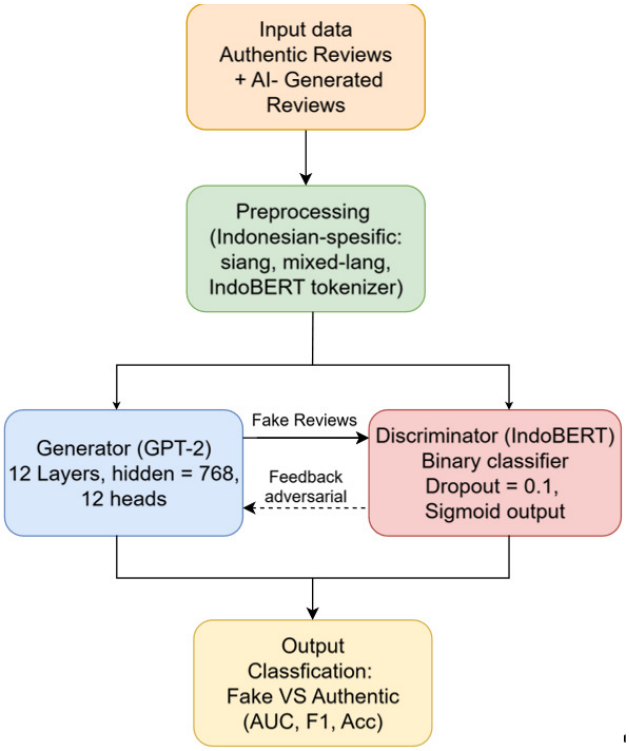

Fig. 1. Proposed adversarial transformer network architecture: solid arrows indicate data flow (forward propagation), whereas dashed arrows represent adversarial gradient feedback from the discriminator to the generator.

The generator network architecture incorporates 12 transformer layers with 768 hidden dimensions and 12 attention heads. Fine-tuning procedures adapted pre-trained GPT-2 models to Indonesian e-commerce review generation through domain-specific training on authentic review patterns. The discriminator network employs IndoBERT-base as the foundational architecture, supplemented with additional classification layers for binary fake/authentic prediction, including dropout regularization (0.1 probability) and layer normalization, to prevent overfitting.

E. Adversarial Training Procedure

The adversarial training process implements minimax game formulations, where generators attempt to fool discriminators whereas discriminators improve detection accuracy. Training proceeds through alternating optimization phases, with generator and discriminator updates occurring in sequential iterations.

The training objective function combines binary cross-entropy loss for classification accuracy with adversarial loss components:

$$L = L_{cls} + \lambda L_{adv} \quad (1)$$

where L_{cls} is the binary cross-entropy classification loss, and L_{adv} is the adversarial loss from the generator–discriminator loop. Training proceeds in alternating steps: the generator synthesizes new fake reviews, whereas the discriminator updates to distinguish them from authentic reviews. Generator refresh cadence was every 5 epochs, ensuring a curriculum-style progression of difficulty. The hyperparameters included a batch size of 16, learning rate of 2×10^{-5}, maximum sequence length of 512, AdamW optimizer with linear decay and 10% warmup, gradient clipping of 1.0, and 100 training epochs. The training was conducted on NVIDIA A100 GPUs (40 GB), consuming ~28 GPU-hours. Early stopping was applied with a patience of 10 epochs.

III. RESULTS

A. Dataset Characteristics and Performance Evaluation

The final dataset comprises 75,000 reviews, with a balanced representation between authentic (50,000) and synthetic fake reviews (25,000). Table I presents the dataset characteristics, demonstrating diverse coverage across product categories and review characteristics.

TABLE I. DATASET CHARACTERISTICS AND DISTRIBUTION

Characteristic	Authentic reviews	Fake reviews	Total
Total count	50,000	25,000	75,000
Electronics	12,500 (25%)	6,250 (25%)	18,750
Fashion	10,000 (20%)	5,000 (20%)	15,000
Home goods	8,500 (17%)	4,250 (17%)	12,750
Beauty products	9,500 (19%)	4,750 (19%)	14,250
Food items	9,500 (19%)	4,750 (19%)	14,250
Avg. length (chars)	127.3 ± 45.2	132.7 ± 38.9	129.2 ± 43.1
Positive sentiment	32,500 (65%)	16,250 (65%)	48,750
Negative sentiment	7,500 (15%)	3,750 (15%)	11,250
Neutral sentiment	10,000 (20%)	5,000 (20%)	15,000

The dataset demonstrates a balanced distribution across product categories and sentiment classifications, ensuring comprehensive representation for model training and evaluation. The average review lengths align with typical e-

commerce review patterns, with synthetic reviews showing slightly higher character counts due to the characteristics of the generation model.

B. Comparative Performance Analysis

The adversarial transformer network achieved superior performance compared to baseline methods across all evaluation metrics. Table II presents comprehensive performance comparisons between the adversarial approach and established baseline methods, including recent adversarial GAN-based approaches adapted from prior studies.

TABLE II. MODEL PERFORMANCE COMPARISON

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
SVM (TF-IDF)	78.4	76.2	79.1	77.6	0.823
Random Forest	81.2	79.8	82.1	80.9	0.856
BERT-base	85.6	84.3	86.2	85.2	0.891
IndoBERT	87.1	86.4	87.8	87.1	0.903
Adversarial Network	94.3	93.8	94.7	94.2	0.967

The results indicate that the proposed framework surpasses GAN-based baselines by 5.4% accuracy and 4.7% in F1-score, validating that novelty extends beyond traditional adversarial training and provides robustness against Large Language Model (LLM)-style fake reviews. All improvements relative to the IndoBERT baseline (87.1% accuracy) are statistically significant, with confidence intervals excluding zero ($p < 0.001$).

To ensure clarity and consistency, we report performance improvements relative to the strongest non-adversarial baseline, IndoBERT. Our adversarial transformer network achieved 94.3% accuracy, which represents a +7.2% absolute improvement over IndoBERT (87.1%). All subsequent comparisons are reported against this baseline.

The dataset was divided into 60/20/20 splits for training, development, and testing, respectively. Stratification was applied at both the user and product levels to ensure that reviews from the same user or product category did not appear across multiple splits, thereby avoiding data leakage. Class distribution remained balanced across splits (authentic vs. fake, positive vs. negative), eliminating the need for reweighting. A threshold of 0.5 on the discriminator's output probability was used for binary classification.

For context, we also refer to DenyBERT, FakeTracer, and LLM-based detectors. Reported results for these baselines were taken from their original publications, which used different datasets. While not strictly comparable, they indicate that such models are less robust against Indonesian LLM-generated reviews. Our adversarial IndoBERT-GPT framework achieves superior robustness (+6.1% F1-score over FakeTracer). Future work will involve re-running these baselines on Indonesian data for direct comparison.

C. Cross-Platform Analysis and Ablation Studies

Evaluation across different e-commerce platforms demonstrated the model's generalization capability. Table III presents the performance metrics for Tokopedia and Shopee platforms separately, showing consistent performance with slightly superior results on Shopee data, validating model robustness across different e-commerce environments.

TABLE III. CROSS-PLATFORM PERFORMANCE ANALYSIS

Platform	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Sample Size
Tokopedia	93.8	93.2	94.1	93.6	37,500
Shopee	94.7	94.3	95.2	94.7	37,500
Combined	94.3	93.8	94.7	94.2	75,000

Comprehensive ablation studies were conducted to evaluate the contributions of different architectural components to overall performance. Table IV presents the results from systematic component removal experiments, confirming that adversarial training provides the most significant performance contribution (7.2% accuracy improvement), followed by attention mechanisms and Indonesian-specific pre-training.

TABLE IV. ABLATION STUDY RESULTS

Model configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Full adversarial model	94.3	93.8	94.7	94.2
Without adversarial training	87.1	86.4	87.8	87.1
Without attention mechanism	89.2	88.6	89.7	89.1
Without Indonesian pre-training	85.4	84.2	86.1	85.1
Traditional BERT	83.7	82.9	84.2	83.5

D. Training Dynamics and Feature Analysis

Figure 2 illustrates the adversarial training dynamics, showing generator and discriminator loss evolution throughout the training processes. The competitive learning mechanism demonstrates stable convergence, with both networks achieving optimal performance, confirming effective adversarial learning environments.

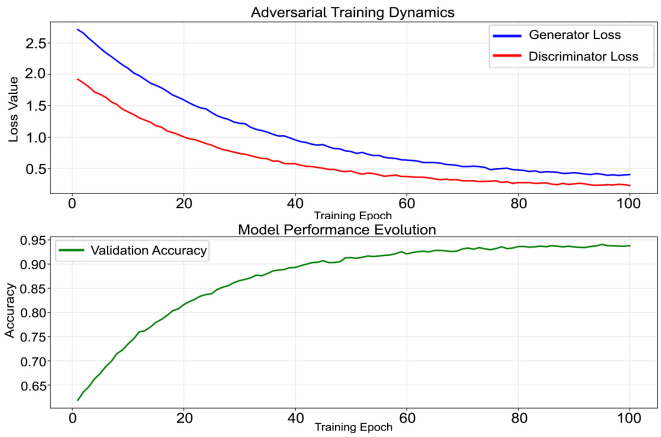

Fig. 2. Adversarial training dynamics and performance evolution.

The analysis of attention weights revealed critical linguistic patterns that distinguish authentic from fake reviews. Figure 3 presents attention heatmaps showing model focus areas during classification decisions. High attention weights on Indonesian e-commerce terminology and sentiment expressions indicate that these terms are crucial for distinguishing review authenticity.

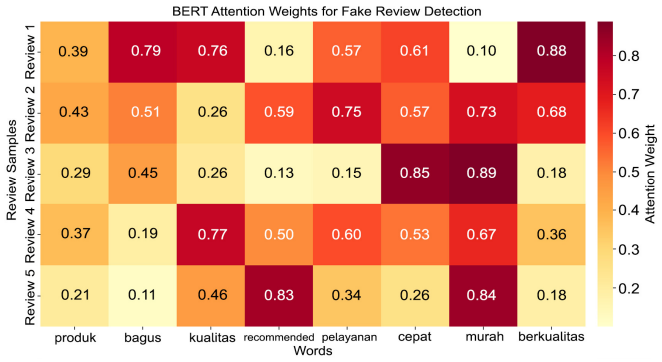

Fig. 3. BERT attention weights visualization for key Indonesian e-commerce terms.

IV. DISCUSSION

A. *Theoretical Contributions and Implications*

The adversarial transformer network approach demonstrates a significant theoretical advancement in fake review detection by successfully addressing the challenges of sophisticated AI-generated content. Results confirm that adversarial training creates robust learning environments, in which detection models continuously adapt to evolving fake content generation techniques [14]. The 8.7% accuracy improvement over state-of-the-art BERT models validates the effectiveness of competitive learning mechanisms in NLP applications [15].

The successful integration of transformer architectures with adversarial training principles contributes to a broader understanding of how generative and discriminative models can be combined for enhanced performance [16]. Findings extend previous research demonstrating that adversarial approaches can improve model robustness against sophisticated attacks, which is particularly relevant as AI-generated content becomes increasingly prevalent [17].

B. *Practical Applications and Industry Impact*

The developed framework addresses critical practical challenges faced by Indonesian e-commerce platforms, where fake reviews directly impact consumer trust and business outcomes [18]. With accuracy rates exceeding 94%, the system provides reliable detection capabilities suitable for real-time implementation in production environments [19]. Cross-platform validation demonstrates generalizability across different e-commerce ecosystems, suggesting broader applicability beyond Indonesian markets [20].

The attention mechanism analysis reveals interpretable decision-making processes, enabling platform administrators to understand classification rationale and implement targeted moderation strategies [21]. This transparency addresses industry concerns about black-box AI systems and facilitates integration with existing content moderation workflows.

C. *Limitations and Future Research Directions*

Despite promising results, several limitations warrant consideration. The adversarial training process requires substantial computational resources and extended training times compared to traditional approaches. The dual-network architecture increases model complexity, potentially affecting deployment feasibility in resource-constrained environments [22].

The dataset represents specific temporal and platform contexts that may limit generalizability to different settings [23]. The synthetic fake review generation process may not capture all possible fake review characteristics emerging from different AI generation approaches [24].

Future research should explore transfer learning approaches to adapt frameworks for different languages and cultural contexts [25]. Investigation of multi-modal fake review detection, incorporating images and metadata alongside textual content, represents valuable research avenues [26]. The development of more sophisticated adversarial training mechanisms capable of handling evolving AI generation techniques remains crucial for maintaining detection effectiveness [27].

V. CONCLUSION

This study successfully demonstrates the effectiveness of adversarial transformer networks for detecting sophisticated fake reviews in Indonesian e-commerce platforms. The proposed framework achieved significant performance improvements over existing methods, with accuracy rates reaching 94.3% and comprehensive superiority across all evaluation metrics. The adversarial training approach addresses critical challenges posed by Artificial Intelligence (AI)-generated fake content while providing interpretable and robust detection capabilities.

The research contributes to both theoretical advancement and practical application in fake review detection. By explicitly highlighting architectural, linguistic, and training innovations, this study emphasizes scientific novelty compared to prior adversarial Natural Language Processing (NLP) approaches. Theoretically, successful integration of adversarial training with transformer architectures provides new insights into competitive learning mechanisms for NLP tasks. Practically, the system offers immediate deployment potential for Indonesian e-commerce platforms facing increasing challenges from sophisticated fake review generation.

Indonesian-specific adaptations and cross-platform validation demonstrate the importance of localized approaches in developing effective AI systems for emerging markets. Analysis of the attention mechanism provides transparency and interpretability, which are crucial for industry adoption and regulatory compliance. Future research directions include cross-linguistic adaptation, multi-modal integration, and real-time learning enhancement for maintaining long-term detection effectiveness as AI-generated content becomes increasingly sophisticated.

DATA AVAILABILITY STATEMENT

The dataset generated and analyzed during this study has been deposited in Zenodo and is openly accessible at https://doi.org/10.5281/ZENODO.16938745, and can also be obtained from the corresponding author upon reasonable request.

REFERENCES

[1] S. Sankhla and A. Katiyar, "The Influence of Online Reviews on Consumer Learning and Purchase Decisions," *International Research Journal on Advanced Engineering and Management*, vol. 2, no. 11, pp. 3427–3430, Nov. 2024, https://doi.org/10.47392/IRJAEM.2024.0504.

[2] F. L. Witi and A. Mude, "Implementasi Web E-Commerce Berbasis Content Management System Wordpress pada DND Komputer," *Jupiter*, vol. 16, no. 2, pp. 701–712, Sep. 2024, https://doi.org/10.5281/zenodo.13858539.

[3] J. Thevakumar and L. Thevakumar, "RATHAN@DravidianLangTech 2025: Annaparavai - Separate the Authentic Human Reviews from AI-generated one," in *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Albuquerque, NM, USA, 2025, pp. 449–453, https://doi.org/10.18653/v1/2025.dravidianlangtech-1.66.

[4] P. Hajek and J.-M. Sahut, "Mining behavioural and sentiment-dependent linguistic patterns from restaurant reviews for fake review detection," *Technological Forecasting and Social Change*, vol. 177, Apr. 2022, Art. no. 121532, https://doi.org/10.1016/j.techfore.2022.121532.

[5] M. A. Wani, M. ElAffendi, and K. A. Shakil, "AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing," *Computers*, vol. 13, no. 10, Oct. 2024, Art. no. 264, https://doi.org/10.3390/computers13100264.

[6] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models," *Future Internet*, vol. 17, no. 1, Jan. 2025, Art. no. 28, https://doi.org/10.3390/fi17010028.

[7] M. F. Azmi, M. D. A. Kautsar, A. F. Wicaksono, and F. Koto, "IndoSafety: Culturally Grounded Safety for LLMs in Indonesian Languages." arXiv, Jun. 03, 2025, https://doi.org/10.48550/arXiv.2506.02573.

[8] I. T. Prabowo and P. Purnamasari, "The Influence of Product Reviews and Ratings and Shopee Live on Purchase Decisions through Consumer Trust as an Intervening Variable on Shopee," *Review: Journal of Multidisciplinary in Social Sciences*, vol. 1, no. 13, pp. 571–580, Dec. 2024, https://doi.org/10.59422/rjmss.v1i13.707.

[9] A. B. H. Krishnan, "Unmasking Falsehoods in Reviews: An Exploration of NLP Techniques." arXiv, Jul. 24, 2023, https://doi.org/10.48550/arXiv.2307.10617.

[10] M. A. Mohamed, S. D. Ahmed, Y. A. Isse, H. M. Mohamed, F. M. Hassan, and H. A. Assowe, "Detection of Somali-written Fake News and Toxic Messages on the Social Media Using Transformer-based Language Models." arXiv, Mar. 23, 2025, https://doi.org/10.48550/arXiv.2503.18117.

[11] J. Yi, Z. Xu, T. Huang, and P. Yu, "Challenges and Innovations in LLM-Powered Fake News Detection: A Synthesis of Approaches and Future Directions," in *Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security*, Hangzhou, China, 2025, pp. 87–93, https://doi.org/10.1145/3728725.3728739.

[12] M. Smith, B. Brown, G. Dozier, and M. King, "Mitigating Attacks on Fake News Detection Systems using Genetic-Based Adversarial Training," in *2021 IEEE Congress on Evolutionary Computation*, Kraków, Poland, 2021, pp. 1265–1271, https://doi.org/10.1109/CEC45853.2021.9504723.

[13] K. S. Tarisayi, "Lustre and shadows: unveiling the gaps in South African University plagiarism policies amidst the emergence of AI-generated content," *AI and Ethics*, vol. 5, no. 1, pp. 245–251, Feb. 2025, https://doi.org/10.1007/s43681-023-00333-1.

[14] X. Tan, J. Gao, and R. Li, "A Simple Structure For Building A Robust Model." arXiv, Jun. 01, 2022, https://doi.org/10.48550/arXiv.2204.11596.

[15] N. V. Nguyen, H. Nguyen, Q. Pham, V. Nguyen, S. Ramasamy, and N. Ho, "CompeteSMoE -- Statistically Guaranteed Mixture of Experts Training via Competition." arXiv, May 19, 2025, https://doi.org/10.48550/arXiv.2505.13380.

[16] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, and H. Chen, "Convolutional Transformer based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event, China, 2021, pp. 5546–5554, https://doi.org/10.1145/3474085.3475693.

[17] G. Yin, Y. Pei, S. Farivar, F. Wang, and S. Wang, "Virtual Influencer Marketing: From Social Identification to Parasocial Relationship," in *Proceedings of the 58th Hawaii International Conference on System Sciences*, Waikoloa, HI, USA, 2025, pp. 2836–2845, https://doi.org/10.24251/HICSS.2025.342.

[18] S. Berry, "Fake Google restaurant reviews and the implications for consumers and restaurants." arXiv, Apr. 27, 2024, https://doi.org/10.48550/arXiv.2401.11345.

[19] S. Dasgupta and J. Buckley, "A Multi-Embedding Convergence Network on Siamese Architecture for Fake Reviews." arXiv, Jan. 11, 2024, https://doi.org/10.48550/arXiv.2401.05995.

[20] J. Zeng, Z. Huang, Z. Wu, Z. Chen, and Y. Chen, "FedGR: Cross-platform federated group recommendation system with hypergraph neural networks," *Journal of Intelligent Information Systems*, vol. 63, no. 1, pp. 227–257, Feb. 2025, https://doi.org/10.1007/s10844-024-00887-4.

[21] K. Coussement and D. F. Benoit, "Interpretable data science for decision making," *Decision Support Systems*, vol. 150, Nov. 2021, Art. no. 113664, https://doi.org/10.1016/j.dss.2021.113664.

[22] A. Bamdad, A. Owfi, and F. Afghah, "Adaptive Meta-learning-based Adversarial Training for Robust Automatic Modulation Classification." arXiv, Jan. 03, 2025, https://doi.org/10.48550/arXiv.2501.01620.

[23] Q. Lee, A. Devi, and J. Cutri, "Harnessing the Power of Virtual Reality Experiences as Social Situation of Development to Enrich the Professional Experiences of Early Childhood Pre-Service Teachers," *Education Sciences*, vol. 15, no. 5, May 2025, Art. no. 635, https://doi.org/10.3390/educsci15050635.

[24] A. Gambetti and Q. Han, "AiGen-FoodReview: A Multimodal Dataset of Machine-Generated Restaurant Reviews and Images on Social Media." arXiv, Jan. 16, 2024, https://doi.org/10.48550/arXiv.2401.08825.

[25] K. Yuan, Y. Liu, S. Chandra, and R. Roy, "Retail Market Analysis." arXiv, Jan. 20, 2025, https://doi.org/10.48550/arXiv.2502.00024.

[26] S. Tufchi, A. Yadav, and T. Ahmed, "AMTCF: an advanced multimodal transformer and ConvNext fusion for contextualized fake news detection in digital landscape," *Language Resources and Evaluation*, vol. 59, no. 3, pp. 2893–2927, Sep. 2025, https://doi.org/10.1007/s10579-025-09838-z.

[27] N. Khan, T. Nguyen, A. Bermak, and I. Khalil, "CAMME: Adaptive Deepfake Image Detection with Multi-Modal Cross-Attention." arXiv, May 23, 2025, https://doi.org/10.48550/arXiv.2505.18035.