ADINT: Machine Learning-Powered Advertisement Intelligence for Proactive Threat Detection in Nigeria's Digital Ecosystem

Muhammad Nazeer Musa muhammadmusa2502@nda.edu.ng
Department of Cybersecurity
Nigerian Defence Academy,
Kaduna, Nigeria

Habib Mohammed himohamm@ncsu.edu
Department of Computer Science
North Carolina State University,
Raleigh, United States

Nafisat Abdulkadir abdulkadirn@custech.edu.ng
Department of Computer Science
Confluence University of Science and Technology, Osara,
Kogi State, Nigeria

Philip Oshiokhaimhele Odion poodion@nda.edu.ng
Department of Computer Science
Nigerian Defence Academy,
Kaduna, Nigeria

Martins Ekata Irhebhude mirhebhude@nda.edu.ng
Department of Computer Science
Nigerian Defence Academy,
Kaduna, Nigeria

Saifullahi Sadi Shitu shituss@nda.edu.ng
Department of Cybersecurity
Nigerian Defence Academy,
Kaduna, Nigeria

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Digital advertising platforms have become unexpected threat vectors in Nigeria: terrorist groups exploit Facebook ads for recruitment, cybercriminals launder billions through advertising-related fraud, and human traffickers lure victims via deceptive job postings. Yet Nigerian security agencies lack systematic capabilities to monitor advertising content as an intelligence source. This paper presents ADINT, the first machine-learning-based advertisement intelligence framework designed for Nigeria's threat landscape. We construct a domain-informed synthetic dataset of 3,000 advertisements across four categories—benign (54.93%), fraud (22.90%), terrorism (11.70%), and trafficking (10.47%)—incorporating realistic class imbalance, graduated ambiguity, and lexical noise to simulate operational conditions. A six-phase experimental pipeline evaluates four architectures: BERT achieves the highest accuracy (91.33%) with perfect recall on terrorism and trafficking; Random Forest (90.33%) offers a compelling efficiency-accuracy trade-off for resource-constrained

deployment. A proposed two-stage cascade where Random Forest pre-filter plus BERT refinement—is analytically projected to reduce analyst workload by 75–78% and maintain zero false negatives on critical threat classes within the synthetic evaluation environment.

Keywords: ADINT, Threat detection, NLP, BERT, BiLSTM, Nigeria, transfer learning, digital security.

1. Introduction

Global digital advertising expenditure exceeded $740 billion in 2024, with social media accounting for approximately $250 billion Statista Research Department (2024). This infrastructure, built for legitimate marketing, has been systematically exploited by threat actors who leverage micro-targeting and partial anonymity to reach highly specific audiences Zouzou and Varol (2024).

Nigeria occupies a uniquely vulnerable position within this evolving threat landscape. Nigeria occupies a uniquely vulnerable position within this evolving threat landscape. As Africa's most populous nation with over 220 million inhabitants, Nigeria faces multifaceted security challenges: terrorism and insurgency by Boko Haram and ISWAP, sophisticated cybercrime operations recording losses exceeding ₦12 trillion ($8 billion) in 2024 Deloitte Nigeria (2026), and thousands of citizens trafficked annually via fraudulent job postings Joe (2021); Ukhami et al. (2024). Despite these threats, Nigerian intelligence and law enforcement have yet to exploit advertising content as an intelligence source. This paper addresses this critical capability gap by developing, training, and evaluating a machine learning-based Advertisement Intelligence (ADINT) framework on a realistic synthetic dataset, with the aim of enabling proactive, scalable, and automated threat detection.

We present ADINT framework for proactive threat detection in digital advertising, designed for deployment within Nigeria's security architecture. Our principal contributions are: (1) a domain-informed, noise-injected 3,000-instance synthetic dataset with realistic class imbalance; (2) a feature engineering pipeline combining 13 hand-crafted attributes with 1,000 TF-IDF features; (3) systematic comparative evaluation of four ML architectures under standardised experimental conditions; and (4) a deployable two-stage inference system with inter-agency threat routing.

2. Related Works

ADINT as an intelligence discipline was formalised by Vines et al. (2017), who demonstrated that advertising micro-targeting can surveil individuals with accuracy comparable to SIGINT at a fraction of the cost, providing the basis for subsequent unsupervised detection frameworks targeting coordinated inauthentic advertising campaigns Cinus et al. (2025).

Traditional classifiers such as SVMs and Random Forests have demonstrated effectiveness against fraud, spam, and phishing-driven advertisements Abbas et al. (2025), while more recent transformer-based models such as BERT Devlin et al. (2019) have significantly enhanced automatic classification of advertising text Zou et al. (2024). However, nearly all such models are trained on Western-language corpora and transfer poorly to multilingual, culturally specific environments like Nigeria Pakray et al. (2025).

Three strands of Nigeria-specific research motivate this work. Nwankpa document how ISWAP used paid social media advertisements Aina and Ojo (2023) and fake NGO pages to evade algorithmic detection Nwankpa (2025). Deloitte Nigeria catalogue large-scale fraudulent investment advertising Deloitte Nigeria (2026). Latonero establish the humanitarian imperative for automated trafficking-advertisement detection Latonero (2011). Collectively, these studies confirm the operational need for an ADINT system while highlighting the absence of a Nigeria-trained model—the gap we address.

3. Methodology

We adopt the Design Science Research paradigm Hevner et al. (2008), which iterates through problem identification, solution design, development, and evaluation for operational information-systems artifacts. All performance claims are evaluated quantitatively on a held-out test set under controlled, replicable conditions. The pipeline comprises six sequential phases (Figure 1).

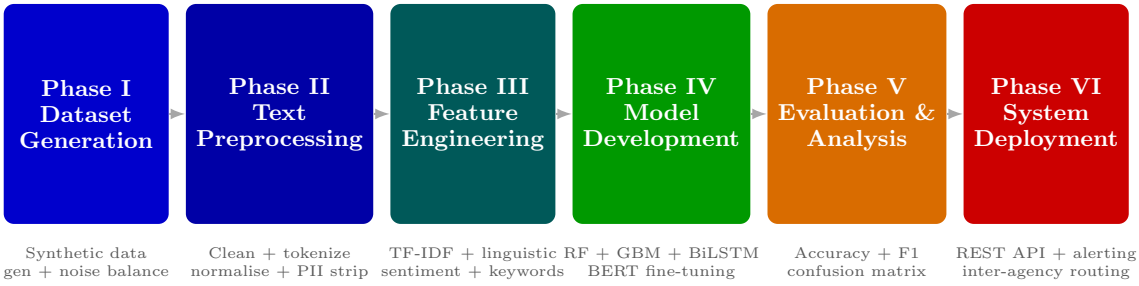

Figure 1: Six-Phase ADINT Research Methodology Workflow

3.1. Phase I: Dataset Design and Synthetic Generation

A fundamental challenge in ADINT research is the near-complete absence of labeled, operational advertisement data: legal constraints, platform access restrictions, and adversary covertness make large-scale authentic collection infeasible. We therefore use domain-informed synthetic generation Lu et al. (2023), an intentional methodological choice that provides full control over class composition, ambiguity profiles, and noise characteristics.

3.1.1. Dataset Design and Threat Category Design

Each category is grounded in three design principles. First, *ecological validity*: vocabulary, grammar, and persuasive strategies are drawn from empirical analyses of real Nigerian malicious advertisements. Second, *realistic class imbalance*: the final class proportions—benign 54.93%, fraud 22.90%, terrorism 11.70%, trafficking 10.47%—mirror the skewed distributions typical of operational intelligence environments (Figure 2). Third, *graduated ambiguity*: each threat category is subdivided into clear positives, ambiguous boundary cases, and edge cases to simulate realistic labeling disagreements and semantic overlap.

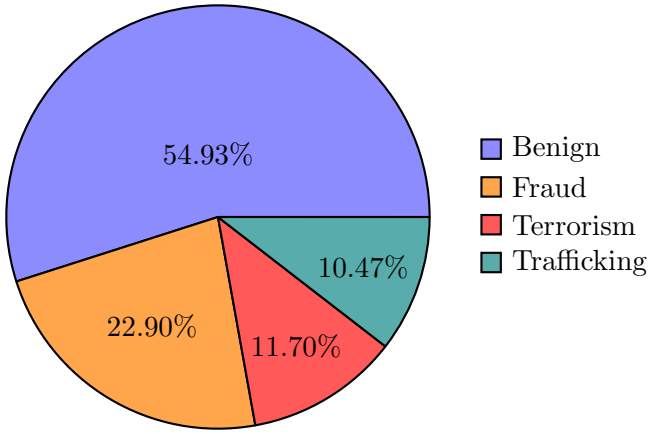

Figure 2: Dataset Composition—3,000 Advertisement Samples with Realistic Class Imbalance

Benign samples include 20% suspicious-looking legitimate examples to stress-test false-positive rates and prevent models from relying on simple lexical cues.

3.1.2. Synthetic Generation Framework

A template-based framework combines category-specific vocabulary pools, agent names, locations, platform metadata, and threat-specific lexicons. Stochastic elements are populated via seeded pseudo-random number generators, ensuring variability and reproducibility. Let the total synthetic dataset be $\mathcal{D}_{syn}$ with $N = 3000$. The generation process is constrained to specific class proportions across $\mathcal{C} = \{c_{\text{benign}}, c_{\text{fraud}}, c_{\text{terrorism}}, c_{\text{trafficking}}\}$:

$$N_{\text{benign}} = 1648, \qquad N_{\text{fraud}} = 687,$$
$$N_{\text{terrorism}} = 351, \qquad N_{\text{trafficking}} = 314.$$

This satisfies $N = \sum_{c \in \mathcal{C}} N_c = 3000$. Following generation, $\mathcal{D}_{syn}$ is uniformly randomly permuted to eliminate order-induced inductive bias.

Stochastic Platform Metadata Assignment Each instance $x_i \in \mathcal{D}_{syn}$ is assigned platform metadata p_i by sampling from a categorical distribution:

$$p_i \sim \text{Categorical}(\mathcal{P}, \Theta)$$

where:

$$\theta_{\text{Facebook}} = 0.40, \quad \theta_{\text{Instagram}} = 0.25, \quad \theta_{\text{WhatsApp}} = 0.15,$$
$$\theta_{\text{Twitter}} = 0.10, \quad \theta_{\text{TikTok}} = 0.05, \quad \theta_{\text{Telegram}} = 0.05.$$

3.1.3. Lexical Noise Injection

To test generalisation under degraded-text conditions typical of social media, two noise mechanisms are applied post-generation.

Character-level perturbation: A noise function $f_{\text{noise}}(x)$ is applied with probability $p_{\text{noise}} = 0.15$:

$$f_{\text{noise}}(x) = \begin{cases} \text{Perturb}(x, S), & \text{with probability } p_{\text{noise}}, \\ x, & \text{with probability } 1 - p_{\text{noise}}. \end{cases}$$

Substitution rules include leet-speak, phonetic abbreviations, and domain misspellings.

Token-Level Stochastic Replacement For each token w_i in the sequence

$$W = (w_1, w_2, \ldots, w_k),$$

a replacement is applied with probability $p_{\text{replace}} = 0.05$:

$$w_i' = \begin{cases} v \sim \mathcal{V}_{\text{neutral}}, & \text{with probability } p_{\text{replace}}, \\ w_i, & \text{with probability } 1 - p_{\text{replace}}. \end{cases}$$

Together, these mechanisms prevent models from relying on brittle syntactic patterns rather than semantic boundaries.

3.2. Phase II: Text Preprocessing

A six-stage deterministic `TextPreprocessor` class normalises surface variation while preserving threat-relevant semantics (Figure 3).

Stage 1—Lowercasing: Reduces vocabulary size and collapses orthographic variants.

Stage 2—URL and Email Removal: Hyperlinks and email addresses are stripped via regular expressions.

Stage 3—Phone Number Stripping: Nigerian phone numbers are removed to prevent spurious correlations.

Stage 4—Special Character Normalisation: Non-alphanumeric characters are removed, except sentence-terminal punctuation.

Stages 5–6—Tokenisation and Lemmatisation: NLTK tokenisation is followed by WordNet lemmatisation and stopword removal.

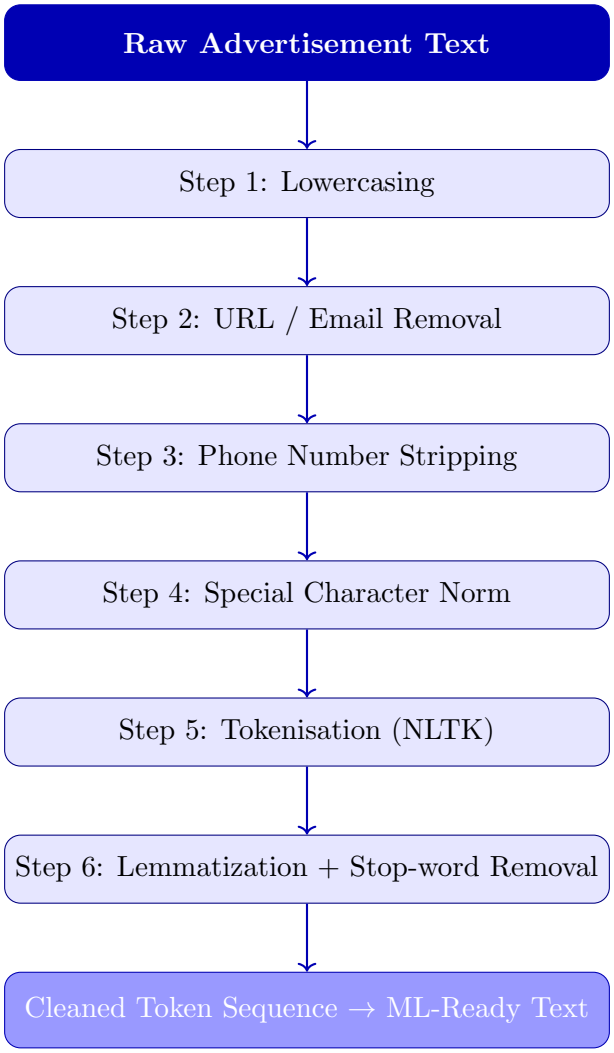

Figure 3: Text Preprocessing Pipeline — Six Sequential Normalisation Stages

3.3. Phase III: Feature Engineering

A 1,013-dimensional feature space is constructed from three groups of hand-crafted features combined with TF-IDF representations (Table 1). Traditional ML models consume this combined feature vector, whereas deep learning models learn their own representations from raw text. The TF-IDF vectoriser is fitted exclusively on the training partition to prevent vocabulary leakage.

3.4. Phase IV: Data Partitioning

$\mathcal{D}_{syn}$ is partitioned via stratified random sampling into training, validation, and held-out test subsets (Figure 4), preserving empirical class distributions across all splits.

Table 1: Feature Engineering Summary

Group	No.	Description	Rationale
Linguistic	8	Character, word, sentence, punctuation, capitalisation, and digit features	Captures urgency and structural markers
Sentiment	2	Polarity and subjectivity scores	Captures emotional tone
Threat keywords	3	Terrorism, fraud, and trafficking lexicon counts	Provides class-specific priors
TF–IDF	1,000	Top unigram, bigram, and trigram features from training data	Captures discriminative n-gram patterns
Total	**1,013**	Sparse feature concatenation	—

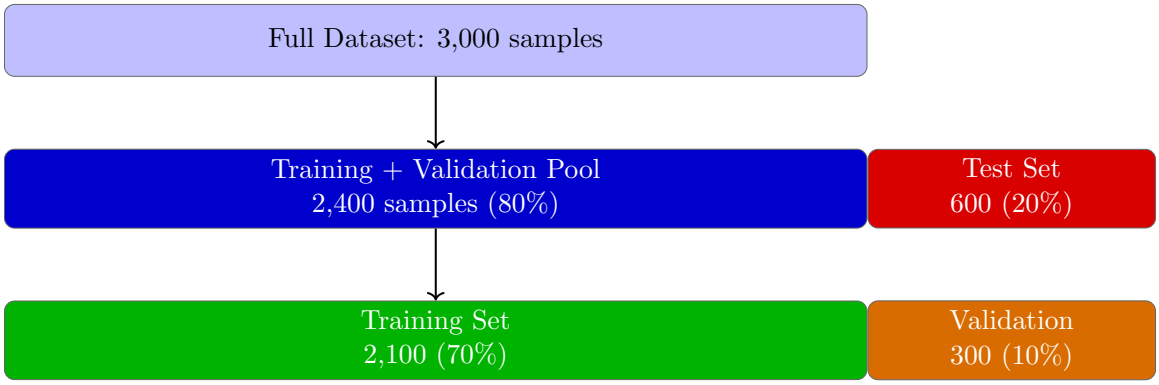

Figure 4: Stratified Dataset Partitioning—70/10/20 Split with Class Preservation

3.5. Phase V: Model Development and Training

Four complementary architectures are evaluated (Figure 5), spanning interpretable ensembles to large pre-trained transformers.

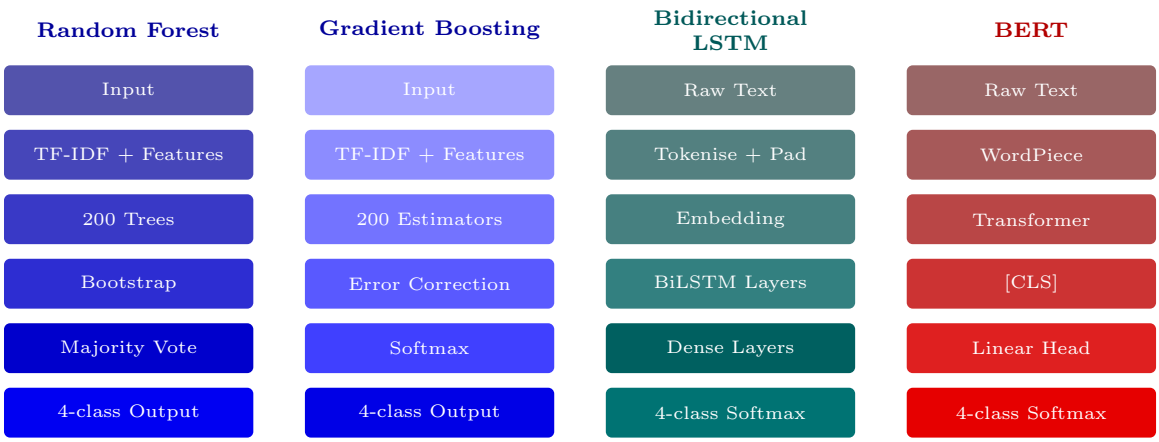

Figure 5: Comparative Summary of the Four ML Model Architectures

3.5.1. Random Forest

200 uncorrelated decision trees are grown using bootstrap aggregation and random feature subsampling; final predictions are determined by majority vote across all trees Breiman (2001).

3.5.2. Gradient Boosting

A sequential ensemble iteratively fits shallow decision trees to the negative gradient of the cross-entropy loss Friedman (2001).

3.5.3. Bidirectional LSTM

The BiLSTM processes sequences in both forward and backward directions, concatenating hidden states at each time step Schuster and Paliwal (1997).

3.5.4. BERT Transformer

`bert-base-uncased` is fine-tuned for sequence classification by appending a linear head to the `[CLS]` token representation Devlin et al. (2019); Vaswani et al. (2017).

Collectively, Random Forest, Gradient Boosting, BiLSTM, and BERT have each demonstrated strong performance in related fraud detection tasks (Abbas et al. (2025)Alzahrani et al. (2025)Taneja et al. (2025). Their deliberate diversity allows us to assess whether deep learning's benefits outweigh its computational costs in Nigerian security applications.

3.6. Phase VI: System Deployment and Prototype Implementation

The deployment architecture rests on layered inference, actionable routing, and human-in-the-loop supervision as shown in Figure 7.

3.6.1. Data Ingestion Layer

In full deployment, a hybrid ingestion layer aggregates advertisements from six platforms via official APIs and transparency databases such as the Meta Ad Library.

3.6.2. Two-Stage Inference Pipeline

The Random Forest classifier screens the full ingested stream as a rapid pre-filter; BERT is then applied exclusively to the flagged subset.

3.6.3. Confidence-Based Alert Routing

Detected threats are stratified into three severity tiers by predictive confidence score (Table 2).

0.45
(a)

(a) BiLSTM Training Dynamics

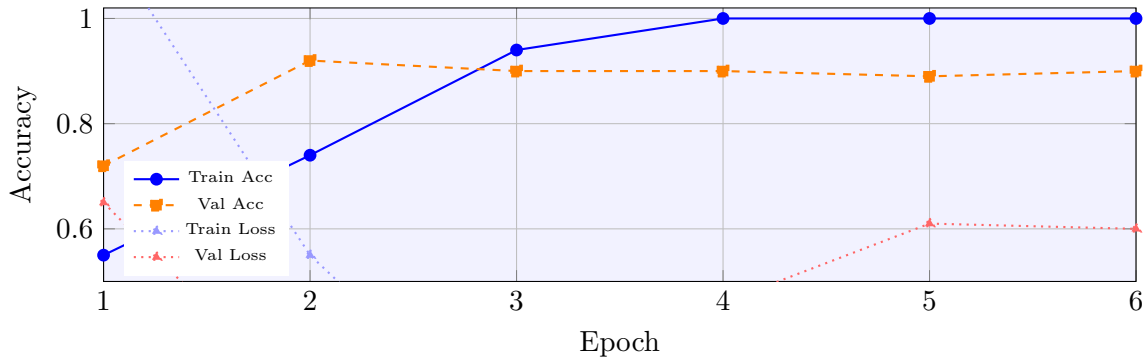

Figure 6: BiLSTM

0.45
(a)

(b) BERT Fine-tuning Dynamics

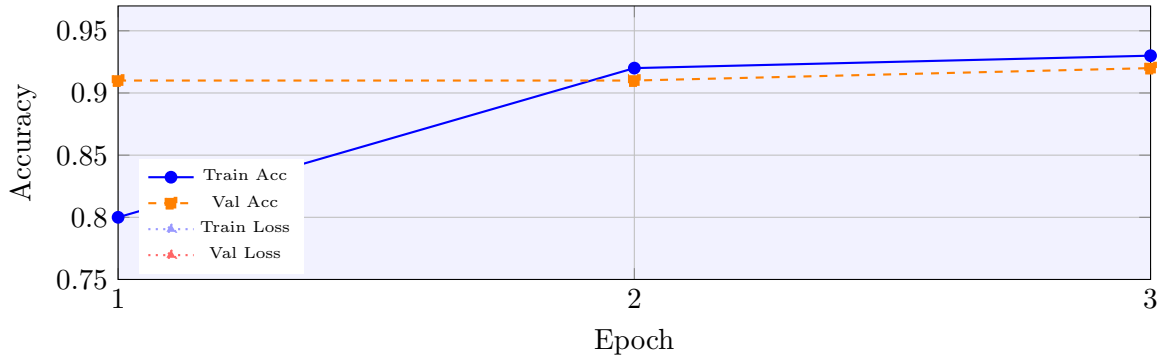

Figure 7: BERT

Figure 8: Training and validation convergence curves for BiLSTM and BERT.

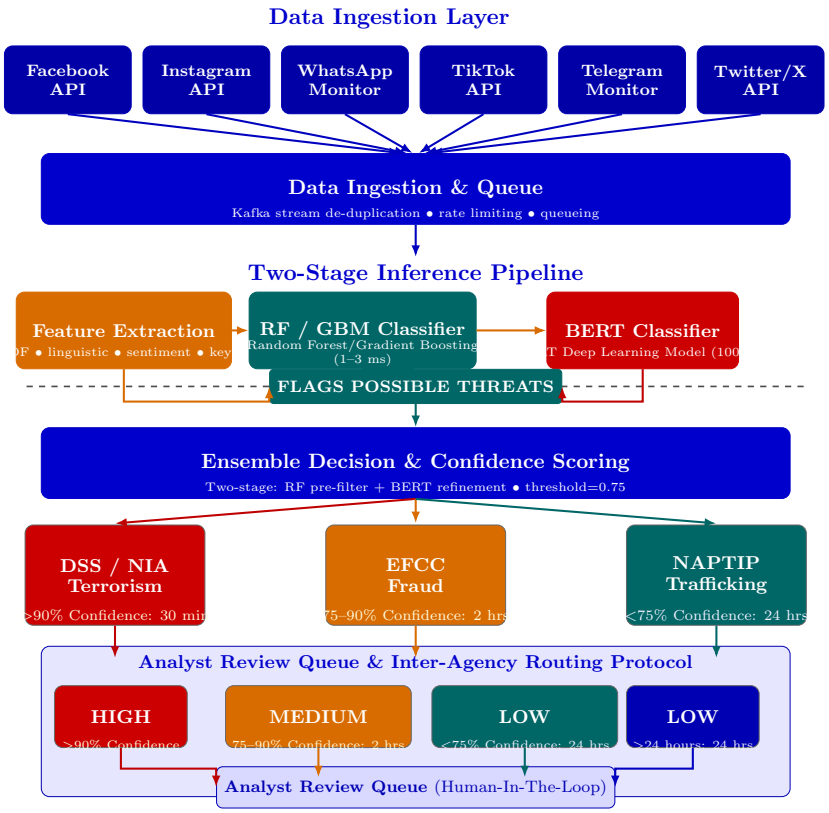

Figure 9: ADINT System Deployment Architecture — Data Ingestion, Inference, and Inter-Agency Routing

Table 2: Confidence-Based Alert Severity and Inter-Agency Routing Protocol

Severity	Confidence	Response SLA	Routing Agency
HIGH	$> 90\%$	30 minutes	DSS/NIA, EFCC, NAPTIP
MEDIUM	75–90%	2 hours	Relevant agency with analyst review flag
LOW	$< 75\%$	24 hours	Analyst review queue—no automatic dispatch

3.6.4. Model Governance and Retraining Protocol

Continuous logging tracks per-class recall against analyst-validated ground truth. If recall for any threat class falls below a defined threshold, the synthetic generation framework can rapidly produce additional training data reflecting new adversarial patterns.

4. Experimental Results

Experiments ran on commodity cloud hardware: NVIDIA T4 GPU for deep models, multi-core CPU for tree-based models; Python 3.10, scikit-learn, PyTorch/TensorFlow.

4.1. Overall Model Performance

All four models were evaluated on the held-out test set of 600 advertisements. Performance is reported across accuracy, weighted F1-score, training time, and inference latency, as summarised in Table 3. The performance spread across all models is a mere 2.50 percentage points (88.83%–91.33%), indicating that the task is amenable to multiple algorithmic approaches and that the engineered feature space provides a highly informative signal for the tree-based models.

Table 3: Comparative Model Performance on 600-Sample Held-out Test Set

Model	Test Acc.	Wtd. F1	Training	Inference
BERT	91.33%	0.91	∼15 min GPU	∼100 ms
Random Forest	90.33%	0.90	∼2 min CPU	<1 ms
Gradient Boosting	89.83%	0.90	∼5 min CPU	<1 ms
Bidirectional LSTM	88.83%	0.89	∼12 min GPU	∼5 ms

4.2. Per-Class Performance Analysis

Confusion matrices in Figure 8 reveal distinct operational profiles with direct deployment implications.

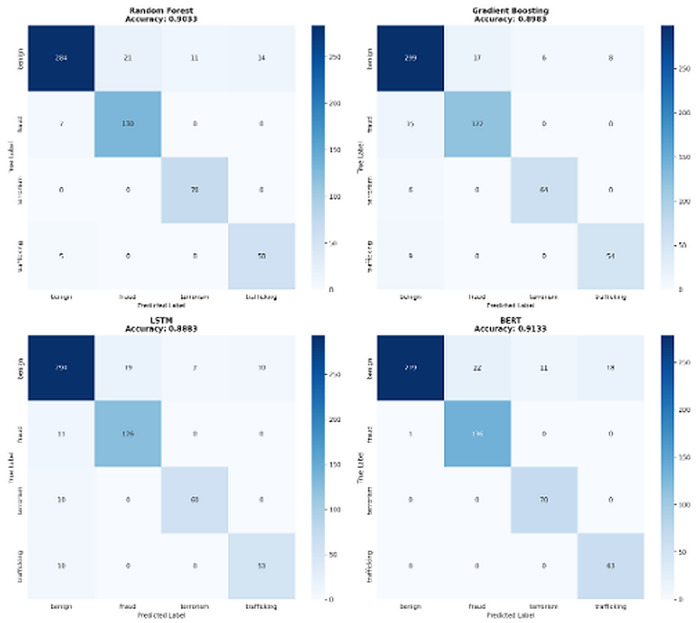

Figure 10: Confusion Matrices

BERT achieves perfect recall on terrorism and trafficking within the synthetic test set at the cost of lower benign recall.

LSTM shows balanced trafficking performance and the second-best fraud recall but trails overall at 88.83% accuracy.

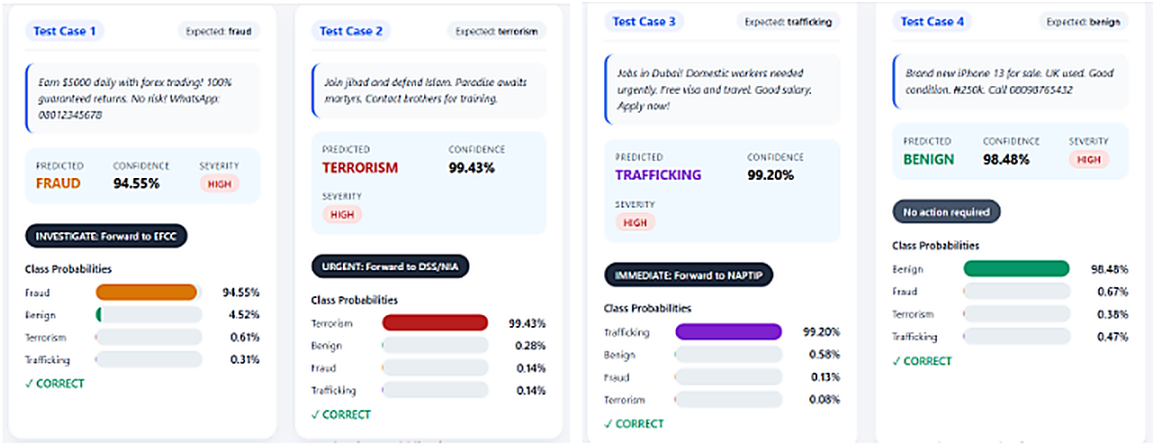

Figure 11: ADINT Prediction System—Test Case Visualizations

Gradient Boosting exhibits the most balanced precision-recall profile with the narrowest inter-class F1 range.

Random Forest matches BERT's perfect terrorism recall with higher benign precision than all other models, producing fewer total false positives.

4.3. Workload Reduction Analysis

Analyst workload reduction is a primary operational metric for ADINT deployment. Prior to automation, security analysts must manually review 100% of flagged advertising content. The ADINT system pre-screens ingested advertisements, escalating only those exceeding the 0.75 confidence threshold. Based on the Random Forest test performance (benign recall of 0.86, flagging approximately 43.2% of all ads), the single-stage system would already reduce the review volume from 100% to 43.2% which is a 56.8% workload reduction. However, the proposed two-stage architecture, in which a lightweight Random Forest pre-filter would screen the full stream before BERT analysis, offers the potential for further workload reductions, followed by BERT analysis of only that flagged subset, the final volume requiring human review could be reduced to approximately 22–25% of the original stream—a projected workload reduction of 75–78%. This cascaded design, which remains a key direction for future implementation, would allow the system to combine Random Forest's efficiency with BERT's semantic depth, enabling analysts to redirect cognitive resources from routine screening to high-priority case analysis, threat pattern research, and inter-agency coordination.

4.4. Qualitative Evaluation

The prototype was tested on representative samples from all four threat categories, correctly classifying each with confidence exceeding 94% as shown in Figure 9.

5. Discussion

5.1. Model Selection for Operational Deployment

The 1-percentage-point gap between BERT and Random Forest challenges transformer dominance assumptions at moderate dataset scales. RF's sub-millisecond, CPU-only inference aligns with Nigerian security infrastructure realities. BERT's zero false negative rate on critical threat classes makes it a strong candidate where missing a true positive carries unacceptable consequences, though real-world validation remains required.

5.2. Misclassification Patterns

Confusion matrix analysis in Figure 10 identified three systematic misclassification patterns that illuminate the fundamental linguistic challenges of advertisement threat detection. First, benign advertisements were the most frequently misclassified category often flagged as fraud because the aggressive urgency-and-scarcity language of legitimate high-pressure marketing and fraudulent schemes occupies overlapping regions of the feature space. This pattern, documented extensively in deception detection literature Whitty (2013), suggests that surface-level linguistic features are insufficient for disambiguation and that contextual, narrative-level understanding is required. Second, bidirectional confusion between trafficking and fraud arises because both categories employ recruitment-and-monetary-incentive language. Third, terrorism demonstrated the sharpest semantic separation, with minimal cross-category confusion, attributable to the presence of highly distinctive ideological vocabulary that constitutes an exclusive lexical cluster absent from other categories.

5.3. Limitations

Six limitations bound this work: (i) all metrics measure in-distribution synthetic generalisation, not real-world performance; (ii) the English-only dataset excludes Hausa, Yoruba, Igbo, and Pidgin English; (iii) the two-stage cascade is unvalidated as a unified pipeline; (iv) confidence-based routing lacks calibration analysis; (v) static threat lexicons require continuous curation; and (vi) adversarial evasion strategies remain unevaluated.

5.4. Ethical and Civil Liberties Considerations

Automated threat routing raises civil liberties concerns. BERT's 15% benign misclassification rate risks unwarranted scrutiny of legitimate advertisers; human-in-the-loop review mitigates but does not eliminate this. Routing thresholds require post-hoc calibration, and deployment must comply with the Cybercrimes Act 2015 and Nigeria Data Protection Act 2023.

6. Conclusion

This paper presented ADINT, a six-phase machine learning framework for proactive advertisement threat detection in Nigeria. Evaluation across 3,000 synthetic samples demonstrated promising in-distribution performance, with BERT achieving the highest accuracy and zero

false negatives on terrorism and trafficking within the controlled test environment. A proposed RF+BERT cascade is analytically projected to reduce analyst workload by 75–78%; empirical validation on real-world Nigerian content remains a prerequisite for operational deployment.

Data availability: Code and data are available at https://github.com/muhammadmusa2502-nazeer/ADINT.git

??

References

Zainab A. Abbas, Zaid M. Hilal, and Hayder G. Jabbar. Click fraud detection in online advertising: A comparative study of machine learning models. *International Journal of Safety & Security Engineering*, 15(3), 2025.

Folahanmi Aina and John Sunday Ojo. The "webification" of jihadism: Trends in the use of online platforms, before and after attacks by violent extremists in nigeria. Technical report, Global Network on Extremism and Technology (GNET), July 2023. URL https://gnet-research.org/2023/07/04/the-webification-of-jihadism-trends-in-the-use-of-online-platforms-before-and-after-atta

Rahaf A. Alzahrani, Malak Aljabri, and Raniah M. A. Mohammad. Ad click fraud detection using machine learning and deep learning algorithms. *IEEE Access*, 2025.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Federico Cinus, Marco Minici, Luca Luceri, and Emilio Ferrara. Exposing cross-platform coordinated inauthentic activity in the run-up to the 2024 us election. In *Proceedings of the ACM on Web Conference 2025*, pages 541–559, April 2025.

Deloitte Nigeria. Nigeria cybersecurity outlook 2026, 2026. URL https://www.deloitte.com/ng/en/services/consulting/perspectives/nigeria-cybersecurity-outlook-2026.html. Accessed: 2026-03-15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *Management Information Systems Quarterly*, 28(1):6, 2008.

Sarah Chidiebere Joe. *Perspectives of reality: The framing of Boko Haram in legacy and social media*. PhD thesis, University of Huddersfield, 2021.

Mark Latonero. Human trafficking online: The role of social networking sites and online classifieds. *Available at SSRN 2045851*, 2011.

Ying Lu, Liang Chen, Yong Zhang, Meng Shen, Hao Wang, Xiaofeng Wang, and Wei Wei. Machine learning for synthetic data generation: A review. *arXiv preprint*, 2023. arXiv:2302.04062.

Michael Nwankpa. Boko haram 2.0? the evolution of a jihadist group since 2015. *Current Trends in Islamist Ideology*, 36, 2025.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197, 2025.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

Statista Research Department. Distribution of global advertising expenditure worldwide by medium. https://www.statista.com/statistics/269333/distribution-of-global-advertising-expenditure/, 2024. Accessed: 2026-03-15.

Kapil Taneja, Jyoti Vashishtha, and Saroj Ratnoo. Fraud-bert: Transformer based context aware online recruitment fraud detection. *Discover Computing*, 28(1):1–16, 2025.

Emmanuel Idemor Ukhami, Agaba Halidu, and Lauretta Azegbeye Achudume. The role of the national agency for the prohibition of trafficking in persons (naptip) in combating human trafficking in nigeria. *Journal of Political Discourse*, 2(1):118–128, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Paul Vines, Franziska Roesner, and Tadayoshi Kohno. Exploring adint: using ad targeting for surveillance on a budget-or-how alice can buy ads to track bob. In *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society*, pages 153–164, 2017.

Monica T Whitty. The scammers persuasive techniques model: Development of a stage model to explain the online dating romance scam. *British journal of criminology*, 53(4): 665–684, 2013.

Liwei Zou, Zhi He, Chengle Zhou, and Wenbing Zhu. Multi-class multi-label classification of social media texts for typhoon damage assessment: a two-stage model fully integrating the outputs of the hidden layers of bert. *International Journal of Digital Earth*, 17(1): 2348668, 2024.

Yasser Zouzou and Onur Varol. Unsupervised detection of coordinated fake-follower campaigns on social media. *EPJ Data Science*, 13(1):62, 2024.