AI-Based Auction Trust and Fraud Detection Platform: A Predictive System for Manipulated Ratings and Suspicious Transactions in Nigerian Online Marketplaces

Levi Anyehechukwu Nwosu[1], Agbakwuru Aphonsus Onyekachi[2], Bethran Chibuike Amanze[2]
[1] Department of Artificial Intelligence and Robotics,
Faculty of Computing and Cybernetic Sciences,
Abia State University, Uturu, Abia State, Nigeria
[2] Department of Computer Science,
Faculty of Physical Sciences,
Imo State University, Owerri, Imo State, Nigeria
Corresponding author: nwosu.levi@abiastateuniversity.edu.ng
https://doi.org/10.56201/ijcsmt.vol.12.no8.2026.pg1.11

Abstract
Online auction and marketplace platforms such as eBay, Amazon Marketplace, Jumia, and Konga sit at the centre of digital commerce, yet remain exposed to trust-eroding fraud, including fake listings, cloned product images, manipulated ratings, shill bidding, and fraudulent sellers. Within Nigeria, these trust deficits form a substantial barrier to marketplace adoption, a problem worsened by the fact that most commercial fraud-detection tools are trained on Western data and cannot reliably parse Nigerian Pidgin, Yoruba, Igbo, or Hausa code-switching in local listings and reviews. This paper presents the design of TrustBid AI, a predictive artificial-intelligence system built to counter four fraud modalities on Nigerian auction platforms: manipulated ratings, fake reviews, cloned product images, and suspicious bidding activity. The architecture rests on four cooperating components: behavioural modelling, content analysis, network interaction signals, and human feedback loops, fused into a hybrid model that reads text, images, and behavioural sequences together rather than in isolation. Behavioural modelling flags shill bidding and bid shielding through sequence-based anomaly detection; content-analysis models score ratings, reviews, and product photographs for manipulation; network analysis maps buyer–seller relationships to expose coordinated fraud rings; and a moderator-in-the-loop layer lets verified reviewers correct and refine model output. Outputs from the four layers are combined into weighted Seller Trust, Item Trust, and Image Authenticity scores rendered on a buyer-facing mobile interface and a moderator web console, both of which are illustrated with interface mock-ups in this paper. A functional prototype is proposed for evaluation on locally sourced Nigerian marketplace data, governed by the Nigeria Data Protection Regulation and its successor, the Nigeria Data Protection Act 2023, with a companion framework for fraud-intelligence sharing among Nigerian platform operators. The expected outcome is a deployable, language-aware trust layer that improves fraud-detection accuracy, reduces buyer exposure to deceptive content, and strengthens confidence in Nigeria's growing digital economy.

Keywords: *artificial intelligence, online auction platforms, fraud detection, machine learning, manipulated ratings, suspicious transactions, Nigerian datasets.*

1. Introduction

Online auction and marketplace platforms have become one of the fastest-growing channels of retail commerce in the last decade, with global players such as eBay and Amazon Marketplace joined in Sub-Saharan Africa by platforms such as Jumia and Konga. Their growth has, however, been matched by an equally rapid growth in the sophistication of the fraud that runs through them. Sellers list counterfeit or misrepresented goods, images are lifted or manipulated to disguise the true condition of an item, ratings are inflated or deflated by coordinated actors, and bidders collude to drive up prices through shill bidding [1], [2]. Because almost every purchasing decision on these platforms is mediated by a rating, a review, or a product photograph, the integrity of these signals is not a peripheral concern but the foundation on which buyer trust is built. In Nigeria specifically, e-commerce adoption has continued to climb even as consumer confidence in platforms such as Jumia and Konga remains fragile, largely because of recurring exposure to counterfeit products, non-delivery, and misleading listings [3]. Existing commercial fraud-detection tools compound the problem: they are typically trained on English-language, Western-market data and are poorly equipped to parse the Nigerian Pidgin, Yoruba, Igbo, and Hausa code-switching that characterizes everyday listings and reviews on Nigerian platforms [4]. A detection system that cannot read the language its users actually write in will always miss a share of the fraud it is meant to catch.

This paper presents the design of TrustBid AI, an AI-based predictive system for detecting manipulated ratings and suspicious transactions on auction and marketplace platforms operating in the Nigerian context. The system is scoped around four fraud modalities: manipulated ratings, fake reviews, cloned or manipulated product images, and suspicious bidding transactions, and is built as a hybrid architecture that reasons over text, images, and behavioural sequences together, rather than treating each signal in isolation. The remainder of the paper reviews related work (Section 2), states the research gap and objectives (Section 3), describes the proposed system architecture and methodology (Section 4), presents the interface design of the prototype (Section 5), outlines the evaluation strategy (Section 6), discusses ethical and data-governance considerations (Section 7), and closes with expected contributions and directions for future work (Sections 8 and 9).

2. Related Work

2.1 Manipulated Ratings and Fake Review Detection

Fake and manipulated reviews have attracted sustained research attention because of their direct effect on consumer purchasing decisions. Salminen, Kandpal, Kamel, Jung, and Jansen [5] built and evaluated machine-learning classifiers for distinguishing genuine from fabricated product reviews, showing that linguistic and metadata features together outperform either feature set alone. Fadhel, Attia, and Ali [1] proposed a fake-review detection pipeline that combined Harris Hawks optimisation with machine-learning classifiers to improve feature selection and detection accuracy. At the market level, He, Hollenbeck, and Proserpio [2] analysed the economics of fake reviews and demonstrated that reputation inflation through fabricated ratings measurably distorts buyer behaviour, reinforcing the case for automated detection rather than manual moderation alone. None of these studies, however, evaluate their models against Nigerian-English, Pidgin, or code-switched review text, which is the gap this research addresses.

2.2 Shill Bidding and Behavioural Fraud Detection

Shill bidding, where a seller or an affiliated party places bids with no intention of winning simply to inflate the final price, remains one of the hardest auction frauds to detect because it mimics genuine bidding behaviour. Gerritse and van Wesenbeeck [6] revised an existing collusive shill-bidding detection algorithm that represents bidders as nodes in a graph and applies Local Outlier Factor scoring together with Belief Propagation over a Markov Random Field to flag coordinated bidding rings, reporting that the approach converges reliably even though ground-truth accuracy could not be verified on unlabelled commercial data. This behavioural, graph-aware approach to shill-bidding detection directly informs the behavioural-modelling and network-interaction components of the architecture proposed in this paper.

2.3 Cloned and Manipulated Image Detection

Because sellers frequently lift product photographs from competitors or manipulate existing images, image-level verification has become a necessary complement to text-based fraud detection. Jakhar and Borah [7] combined perceptual hashing with deep-learning feature embeddings to detect near-duplicate images with higher accuracy and lower false-positive rates than either technique used alone, a hybrid strategy well suited to catching cloned product photographs across auction listings. This body of work underpins the image-verification layer of the proposed system, which pairs perceptual hashing with convolutional feature matching to flag lifted or manipulated listing photographs.

2.4 Graph-Based and Network-Level Fraud Detection

A parallel stream of research treats fraud as a network problem rather than a per-transaction one. Kim, Choi, and Whang [8] proposed a dynamic relation-attentive graph neural network that aggregates signals across multiple relational contexts, including shared authorship, timing, and rating similarity, to identify fraudulent reviewers on benchmark datasets, improving on earlier relation-aware graph models. Such techniques are directly transferable to the buyer–seller interaction graphs that the network-interaction-signals component of TrustBid AI is designed to analyze, since coordinated rating manipulation and collusive bidding both leave a graph-level fingerprint that is difficult to see from any single account's history alone.

2.5 Trust, Language, and the Nigerian E-Commerce Context

Consumer trust research on African e-commerce consistently identifies fraud exposure and weak buyer protection as the leading barriers to sustained platform adoption [3]. A separate but equally important gap is linguistic: most fraud-detection language models are trained on Standard English and underperform when applied to Nigeria's linguistically diverse listings and reviews. Muhammad et al. [4] addressed this gap directly by releasing NaijaSenti, a large, human-annotated Twitter sentiment corpus covering Hausa, Igbo, Nigerian Pidgin, and Yoruba, including a substantial proportion of code-mixed text, and showed that language-specific and language-adaptive fine-tuning consistently outperform generic multilingual models on these languages. This finding motivates the Nigerian-language-aware text-analysis layer proposed in Section 4 of this paper.

3. Problem Statement and Research Objectives

Three gaps recur across the literature reviewed in Section 2. First, language-blind detection: existing fraud-detection models are trained overwhelmingly on Western, English-language datasets and cannot reliably capture the Nigerian Pidgin, Yoruba, Igbo, and Hausa code-switching that appears routinely in Nigerian listings and reviews [4]. Second, single-modality focus: most published systems examine text, images, or bidding behaviour in isolation, which allows fraud that spans multiple modalities, for example, a manipulated image paired with a coordinated review campaign, to slip through detection [5], [7]. Third, limited local evidence: very few fraud-detection systems have been evaluated against data drawn from Nigerian marketplaces such as Jumia, Konga, or Facebook Marketplace, which limits confidence in their real-world reliability within this market.

The research focus of this study is therefore a predictive AI system that spans four fraud modalities: manipulated ratings, fake reviews, cloned product images, and suspicious bidding transactions, purpose-built for Nigerian auction and marketplace platforms. The specific objectives are to: (i) design a hybrid AI architecture that fuses behavioural, textual, visual, and network signals into a single trust assessment; (ii) develop and evaluate individual model layers for text, image, behavioural, and rating-fraud analysis, tuned where relevant for Nigerian-English and code-switched text; (iii) define a weighted trust-scoring mechanism that expresses model output as interpretable Seller, Item, and Image Authenticity scores; (iv) design buyer-facing and moderator-facing interfaces that surface these scores usefully rather than as an opaque single number; and (v) propose an evaluation and data-governance framework consistent with Nigerian regulatory requirements.

4. System Architecture and Methodology

4.1 Overview of the Hybrid AI System

TrustBid AI is organized around four cooperating components. Behavioural modelling captures suspicious seller and bidder actions, such as bid shielding and shill bidding, using sequence-based anomaly detection built on the graph and outlier-scoring techniques reviewed in Section 2.2 [6]. Content analysis applies machine-learning classifiers to detect manipulated ratings and fake reviews in text, and deep-learning models to identify cloned or manipulated product images, drawing on the perceptual-hashing and embedding techniques discussed in Section 2.3 [7]. Network interaction signals examine buyer–seller relationship graphs to surface coordinated fraud patterns that are invisible at the level of a single transaction, following the relation-aware graph approach of Kim et al. [8]. Human feedback loops allow moderators and verified buyers to validate or correct model output, keeping the system accountable and providing a continuous stream of labelled examples for retraining. These four components are fused into a single hybrid pipeline that combines text, image, and behavioural analytics to separate genuine feedback from sentiment-driven bias and coordinated rating manipulation.

4.2 Research Methodology

The study follows a six-stage, design-and-development methodology summarized in Table 1. Datasets are collected locally from Nigerian marketplaces to avoid the Western-data bias identified in Section 3, then cleaned and anonymized before any model training takes place. Four families of AI models are then developed and validated independently before being integrated into a single prototype consisting of a verification engine, a listing analyzer, a behaviour tracker, and a moderator dashboard, which is evaluated using a combination of quantitative metrics, qualitative user testing, and benchmarking against existing platform safeguards.

Stage	Purpose
Dataset Collection	Gather image, textual listing, behavioural/transactional, and ratings data sourced locally from Nigerian marketplaces.
Preprocessing & Anonymisation	Clean and label the data; remove faces, phone numbers, names, and chat identities prior to model training.

AI Model Development	Train image-verification, text-analysis, behavioural-modelling, and rating-fraud detection models.
Trust-Score Computation	Combine model outputs into weighted Seller, Item, and Image Authenticity trust scores.
System Integration	Assemble the prototype: verification engine, listing analyzer, behaviour tracker, and moderator dashboard.
Evaluation	Assess the system through quantitative metrics, qualitative user testing, and benchmarking against existing platforms.

Table 1. Multi-layered methodological approach adopted for system development.

4.3 AI Model Layers

Table 2 summarizes the candidate models assigned to each of the four fraud modalities. Text analysis relies on transformer-based classifiers and TF-IDF baselines tuned for Nigerian-English and code-switched text, following the language-adaptive approach shown to work well by Muhammad et al. [4]. Image verification combines lightweight and high-accuracy convolutional classifiers with vision-language matching and perceptual hashing, consistent with the hybrid hashing-plus-embedding strategy of Jakhar and Borah [7]. Behavioural modelling uses ensemble tree methods, kernel-based classifiers, and clustering to separate trustworthy sellers from suspicious or fraudulent ones, and rating and review fraud detection applies graph-based collusion detection, anomaly detection, and sequence models to catch false ratings, bot-generated reviews, and rating bursts, building on the graph-neural and Markov-random-field approaches reviewed in Sections 2.2 and 2.4 [6], [8].

Layer	Candidate Models	Task
Text Analysis	Transformer-based text classifiers and TF-IDF baselines tuned for Nigerian-English and code-switched text.	Flag misleading descriptions and exaggerated claims.
Image Verification	Lightweight and high-accuracy CNN classifiers, vision-language matching, and perceptual hashing.	Detect cloned, lifted, or manipulated product images.
Behavioural Modelling	Ensemble tree methods, kernel-based classifiers, and clustering.	Classify sellers as trustworthy, suspicious, or fraudulent.
Rating & Review Fraud	Graph-based collusion detection, anomaly detection, and sequence models.	Catch false ratings, bot-generated reviews, and rating bursts.

Table 2. Candidate AI model layers mapped to the four fraud modalities.

4.4 Trust-Score Computation

Outputs from the four model layers are not exposed to buyers as raw classifier scores; they are combined using a weighted scoring model, with ensemble learning and optional fuzzy logic available for soft scoring, into three interpretable measures on a 0–100% scale: Seller Trust Score, Item Trust Score, and Image Authenticity Score. As shown in Table 3, behavioural signals carry the largest share of the composite weighting, reflecting the finding in Section 2.2 that behavioural and network anomalies are the most reliable early indicators of coordinated fraud, followed by image verification, text and listing quality, and network rating patterns.

Signal Category	Weight
Behavioural signals (bid shielding, shill bidding, seller activity anomalies)	40%
Image verification (cloned or manipulated product images)	30%
Text and listing quality (misleading descriptions, exaggerated claims)	20%
Network rating patterns (collusive or bot-generated reviews)	10%

Table 3. Weighting scheme feeding the composite trust score.

5. Prototype Interface Design

To make the trust-scoring architecture concrete, a set of interface mock-ups was developed for TrustBid AI, spanning a buyer-facing mobile application and a moderator-facing web console. Rather than presenting a single opaque number, every screen breaks the composite trust score down into its constituent signals, so that a buyer or moderator can see which factor, whether seller history, item description, or image authenticity, is driving a given score. Risk states use a consistent colour language across every screen: green for trustworthy, amber for suspicious, and red for fraudulent, reinforcing the same visual grammar throughout the application.

5.1 Buyer Experience: Listing Trust Scores and Image Verification

On the listing-detail screen (Figure 1), Seller Trust, Item Trust, and Image Authenticity scores render as live gauges directly beneath the product listing, giving the buyer a fraud-risk signal before they engage with the seller. When a listing photograph matches an indexed external image above the cloning threshold, the image-verification screen (Figure 2) surfaces the perceptual-hash comparison and lets the buyer report the listing in a single tap, operationalizing the image-verification layer described in Section 4.3.

Figure 1 (left). Listing detail screen with the Seller, Item, and Image Authenticity trust-score panel. Figure 2 (right). Image-verification result showing a perceptual-hash match against an indexed cloned image.

5.2 Trust and Safety Signals: Seller Risk and Review Fraud

The seller behavioural risk profile (Figure 3) plots a trust-score trend over time alongside behavioural flags such as price manipulation, refund disputes, and listing churn, giving moderators evidence rather than a bare score. The companion ratings-and-review screen (Figure 4) tags each review as Genuine, Suspicious Cluster, Bot-Generated, or Rating Burst, with an aggregate flagged-review count displayed alongside the conventional star rating, directly reflecting the rating-fraud model layer described in Section 4.3.

Figure 3 (left). Seller behavioural risk profile with trust-score trend and behavioural flags.
Figure 4 (right). Ratings and review fraud-detection screen with per-review classification.

5.3 Moderator Console

Flagged listings, sellers, and reviews are triaged by human moderators through a dedicated web dashboard (Figure 5), which lets reviewers investigate, dismiss, or escalate every flagged case. This console operationalises the human-feedback-loop component of the architecture (Section 4.1), ensuring that automated flags are always subject to human review before any punitive action is taken against a seller.

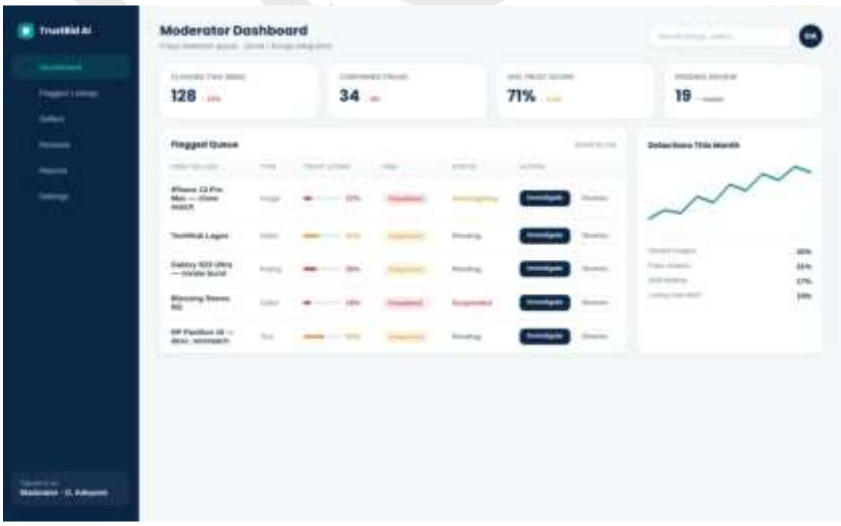

Figure 5. Moderator web console for fraud-case review and management.

5.4 Key Design Principles

Six design principles guided the interface work: real-time trust scoring, so that seller, item, and image scores recomputed continuously and render as gauges a buyer can read at a glance; cloned-image detection, using perceptual hashing and vision-language matching to flag lifted photographs before checkout; rating-fraud flags that classify reviews as genuine, clustered, bot-generated, or burst rather than reducing everything to a star average; Nigerian-language awareness, with text models tuned for Pidgin, Yoruba, Igbo, and Hausa code-switching in listings and reviews; moderator case management through the dedicated web console; and privacy-first data handling, with every interface operating on anonymized data aligned with Nigerian data-protection requirements, discussed further in Section 7.

6. Evaluation Strategy

The prototype is evaluated using a three-part strategy. Quantitative evaluation reports Accuracy, Precision, Recall, and F1-score for each classifier, together with ROC-AUC and confusion-matrix analysis, and Mean Absolute Error for the trust-scoring layer. Qualitative evaluation recruits real Nigerian buyers to test the system directly, comparing AI-driven flags against human judgement and measuring buyer confidence before and after use. Benchmarking compares TrustBid AI's performance against Amazon's rating-verification mechanisms, Jumia and Konga's existing seller-scoring systems, and standard fake-review detectors reported in the literature reviewed in Section 2, so that any improvement can be stated relative to systems already in commercial use.

7. Ethical Considerations and Data Governance

All data collected for model training and evaluation is anonymized before use: faces, phone numbers, names, addresses, and chat identities are removed prior to any processing step. Data handling throughout the project is guided by the Nigeria Data Protection Regulation [9] and its successor framework, the Nigeria Data Protection Act [10]. Anonymisation, informed consent, and data-subject rights are embedded from dataset collection through to trust-score deployment, rather than added retroactively once a prototype already exists. A proposed fraud-intelligence-sharing framework would additionally allow Nigerian platform operators to exchange fraud signals, for example a confirmed shill-bidding ring or a cloned-image source, without exposing the personal data of the individuals involved, extending the collaborative defence model that single-platform systems cannot achieve alone.

8. Expected Outcomes and Contributions

The completed system is expected to deliver five contributions. First, automated flagging of fake or cloned product images at listing time, reducing the manual moderation burden on platform staff. Second, earlier detection of false ratings and scam sellers than is currently possible with single-modality tools. Third, higher buyer trust and confidence in auction and marketplace platforms, measured through the qualitative evaluation strategy described in Section 6. Fourth, improved detection accuracy on Nigerian-specific data relative to systems trained solely on Western datasets, addressing the language gap identified in Section 3. Fifth, a reusable trust-scoring model and interface pattern that other researchers and platform operators can adapt to different African e-commerce contexts beyond the Nigerian market examined here.

9. Conclusion and Future Work

This paper has presented the design of TrustBid AI, a hybrid AI system for detecting manipulated ratings and suspicious transactions on Nigerian online auction and marketplace platforms. By fusing behavioural modelling, content analysis, network interaction signals, and human feedback into a single weighted trust score, and by tuning the text-analysis layer for Nigerian-English and code-switched text, the proposed architecture directly addresses the language-blindness, single-modality focus, and limited local evidence identified as gaps in the existing literature. The interface mock-ups presented in Section 5 demonstrate how a composite trust score can be surfaced usefully to both buyers and moderators without reducing fraud signals to a single, uninterpretable number. Future work will focus on three areas: building and annotating the local Nigerian training datasets described in Section 4.2, implementing and benchmarking the individual model layers against the evaluation strategy in Section 6, and piloting the moderator console with a partner Nigerian marketplace to validate the fraud-intelligence-sharing framework proposed in Section 7 under real operating conditions.

References

[1] Fadhel, Z., Attia, H., & Ali, Y. H. (2022). Optimized and comprehensive fake review detection based on Harris Hawks optimization integrated with machine learning techniques. Journal of Computational Science, 60, 101601.

[2] He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. Marketing Science, 41(5), 896–921. https://doi.org/10.1287/mksc.2022.1353

[3] Chen, Y. (2025). Trust models and data privacy in e-commerce: A case study of Jumia Kenya. Academy of Entrepreneurship Journal, 31(2), 1–12.

[4] Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., Choudhury, M., Emezue, C. C., Abdullahi, S. S., Aremu, A., Jeorge, A., & Brazdil, P. (2022). NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 590–602). European Language Resources Association.

[5] Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, Article 102771. https://doi.org/10.1016/j.jretconser.2021.102771

[6] Gerritse, L. A., & van Wesenbeeck, C. F. A. (2022). Detecting collusive shill bidding in commercial online auctions. Computational Economics, 63, 1–20. https://doi.org/10.1007/s10614-022-10326-7

[7] Jakhar, Y., & Borah, M. (2025). Effective near-duplicate image detection using perceptual hashing and deep learning. Information Processing & Management, 62(4), Article 104086. https://doi.org/10.1016/j.ipm.2025.104086

[8] Kim, H., Choi, J., & Whang, J. J. (2023). Dynamic relation-attentive graph neural networks for fraud detection (arXiv:2310.04171). arXiv. https://arxiv.org/abs/2310.04171

[9] National Information Technology Development Agency. (2019). Nigeria Data Protection Regulation (NDPR). NITDA.

[10] Nigeria Data Protection Commission. (2023). Nigeria Data Protection Act (NDPA).