Journal of Intelligent Decision Support System (IDSS) 8 (4) (2025), pp. 218-236

Published by: Institute of Computer Science (IOCS)

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www.idss.iocspublisher.org

Strategy for preventing human trafficking through verification of online job vacancies in Indonesia

Arga Husein Passu Beta[1], H.A. Danang Rimbawa[2], Aulia Khamas Heikhmakhtiar[3]

[123] *Department of Cyber Defense Technology, Republic of Indonesia Defense University, Bogor, Indonesia*

Article Info

Article history:

Received Oct 9, 2025
Revised Oct 20, 2025
Accepted Nov 11, 2025

Keywords:

Human Trafficking;
Logistic Regression;
Random Forest;
Scam Job Posting;
Selective AI.

ABSTRACT

This study addresses the rise of online job ads used to recruit victims of human trafficking (TPPO). We propose a practical screening approach that combines automated checks with human moderation. The goal is not to prove crimes, but to prioritize high-risk ads for fast review and referral. Using a public dataset of 500 job postings (fake_job_postings_500), we clean the text and basic metadata, extract simple text features (TF–IDF), and add light verification signals (e.g., contact and firm consistency). We then train two models in a leakage-safe pipeline: calibrated Logistic Regression (LR-cal) and Random Forest (RF). Performance is evaluated with standard accuracy measures ROC-AUC, PR-AUC, F1 plus calibration (how well risk scores match reality) and triage metrics that reflect real operations: precision for the highest-risk group, recall for all medium-and-above risk, and the share of ads moderators must review. Results show LR-cal is accurate and well-calibrated (5-fold means: ROC-AUC 0.993, PR-AUC 0.986, F1 0.934). In triage with thresholds T_high = 0.80 and T_med = 0.50, LR-cal yields Precision@High = 1.00 and Recall@≥Med=0.925 with ~34% of ads needing review. RF reaches near-ceiling accuracy (1.00/1.00 at ~35.3% workload) but requires careful calibration and leakage auditing. Practical contribution: AI-assisted, risk-based gatekeeping can reduce exposure to Human Trafficking or TPPO at the source. We recommend: (1) adopting calibrated models with adjustable thresholds; (2) standard operating procedures (SOPs) for cross-platform verification, including Know Your Customer (KYC) and Open-Source Intelligence (OSINT) checks; and (3) direct integration with official reporting channels to escalate flagged ads swiftly.

Corresponding Author:

Arga Husein Passu Beta,
Department of Cyber Defense Technology,
Republic of Indonesia Defense University,
IPSC Sentul Area, Sukahati, Bogor, West Java 16810.
Email: argahussein98@gmail.com

Introduction

Human trafficking (TPPO) is increasingly mediated by technology, including through fake online job postings that lure victims into scam centers. Recent studies caution us to be wary of claims to "expose TPPO" if they rely solely on indicators that have not been strongly validated due to weak ground truth, subjective indicators, and perpetrators' rapidly adapting tactics. In practice, screening risky job advertisements remains necessary as an upstream (preventive) measure, but it is not sufficient for legal

determination without additional evidence and a valid enforcement process. (Giommoni, 2024). Indonesian law defines human trafficking as "the act of recruiting, transporting, harboring, sending, transferring, or receiving for the purpose of exploitation" both domestically and internationally.

Based on IOM operational monitoring, the top five destination countries for forced criminality schemes in scam centers are Lao PDR, Cambodia, Myanmar, the Philippines, and Malaysia, while the top five countries of origin are Vietnam, Indonesia, Kenya, Lao PDR, and India. IOM estimates that tens of thousands of victims are forced to work in online scam centers, mainly located in Myanmar, Cambodia, and Lao PDR, although the exact number is difficult to determine due to the closed nature of the operations .

The profile of victims handled by IOM shows that in 2022, 36% (296 out of 815 cases) were related to exploitation for forced criminality in Southeast Asian "scam" centers; 68% were male, 32% were female, 97% were aged ≥18 years. Victims tend to be young adults (18–35 years old) multilingual, and tech-savvy, with post-COVID economic pressures widening the vulnerability of middle-class workers (IOM UN, 2023). In the context of cross-border employment, similar patterns emerge in risky labor migration practice-based reports in Indonesia highlight a surge in human trafficking during the pandemic and examples of exploitation of Indonesian seafarers on Chineseflagged ships (including cases of seafarers' bodies being dumped at sea), reflecting weak protection and information asymmetry for migrant workers. In the online realm, fake job postings and cross-platform recruitment advertisements have become the main recruitment vectors. The scale of this problem is confirmed by (Vuyyuru, 2024), who cites UNODC data that around 31% of trafficking victims in 2024 were initially targeted through digital platforms, noting a 47% increase in digital recruitment methods between 2019 and 2024. The journal also highlights a key challenge, namely the ability of perpetrators to continuously adapt; their digital strategies are estimated to change on average every 2.5 months, which means that conventional detection methods quickly become obsolete.

An international scoping study of migrant job advertisements found that 98.4% of advertisements contained at least one risk indicator, suggesting that some commonly used indicators may only reflect labor market characteristics rather than evidence of TPPO. Their usefulness for large-scale job advertisement risk assessment tends to be limited (Kleinberg, 2020). A global report from the UNODC confirms a shift in patterns and increased vulnerability in the wake of the pandemic (UNODC, 2022). Within Indonesia, anti-trafficking, migrant-worker protection, and personal-data safeguards already exist in law and policy, and public portals for job placement operate alongside private platforms. However, the verification of online vacancies remains challenging: postings proliferate on non-official channels, pre-publication screening is limited, reporting and enforcement are fragmented across institutions, cross-platform evidence-sharing is difficult to operationalize under privacy and purpose-limitation constraints, and perpetrators rapidly adapt their messaging. These factors motivate a verification approach that is both compliant with national regulation and practically integrable with government and platform workflows.

Therefore, advertisements need to be "screened" for suspicion, but additional information is needed to confirm a link to TPPO. In line with this, the latest methodological review warns that detecting TPPO "based on online advertisements alone" is unrealistic due to weak ground truth, the subjectivity of indicators, and the risk of overclaiming. Nevertheless, computational approaches remain promising for large-scale initial screening with strict ethical and methodological caveats. Multi-input deep learning studies on online advertisements show improved precision in high-risk subsets, which is useful for prioritizing law enforcement investigations; however, the authors emphasize the need for up-todate data and validated labeling . In the context of labor recruitment, another study assessed that "labor TPPO indicators" are useful in handling individual cases but "their usefulness for large-scale job advertisement risk assessment tends to be limited," requiring a combination of automation and human-inthe-loop to improve accuracy and reduce bias.

Indonesia's legal framework (IOM UN, 2023) provides a basis for criminalizing recruiters who exploit Indonesian citizens at home and abroad, including significant criminal penalties and fines. This case highlights broader challenges related to human security in navigable spaces (Alvarez Moreno, 2021) . This serves as a normative basis for upstream prevention interventions, namely at the stage of

verifying online job vacancies. However, in terms of victim protection, policy findings highlight practices that are "not yet synchronized," particularly unclear restitution mechanisms and difficulties in proving losses, indicating the need for an effective integrated reporting and service system. Research Gap. Despite growing detection efforts, (i) commonly cited job-fraud indicators have not been rigorously validated on Indonesian-language corpora and labor-market context, risking false positives/negatives; (ii) there is no audited, human-in-the-loop verification pipeline tailored to Indonesian content and interoperable with public portals and enforcement pathways; (iii) privacy-preserving data governance for user-generated content is under-specified relative to national data-protection requirements; and (iv) evaluations seldom account for risk-sensitive metrics and concept drift, even though adversarial messaging evolves quickly.

Based on these gaps, this study focuses on prevention strategies through the verification of online job vacancies in Indonesia. In practical terms, researchers designed an Artificial Intelligence (AI) based verification pipeline for screening job advertisements, AI as a one-stop point (Kusumowijoyo et al., 2023). Thus, this research contributes to upstream prevention: not by identifying TPPO definitively from advertisements, but by reducing recruitment opportunities through verification-based gatekeeping, strengthening risk literacy, and providing early warnings integrated with the law enforcement and victim protection ecosystem. This approach is in line with the methodological warning that "we must be cautious of studies that claim to 'uncover TPPO' when in fact they only apply untested indicators," so that the function of AI is emphasized for screening and prioritization, not legal determination.

Indonesia law No. 21 of 2007 defines human trafficking as a series of acts of recruitment, transportation, harboring, transfer, or receipt of a person with elements of coercion/deception/abuse of a position of vulnerability for the purpose of exploitation, both transnational and domestic. Indonesia law of No. 18 of 2007 has established laws on the Protection of Indonesian Migrant Workers regulating placement and protection. Presidential Regulation No. 19/2023 establishes the 2020–2024 National Action Plan for the Prevention and Handling of Trafficking in Persons and is reinforced by the renewal of the structure of the Task Force on Trafficking in Persons through Presidential Regulation No. 49/2023; while Law No. 27/2022 on Personal Data Protection requires compliance with data processing in automatic detection systems. At the service level, the Ministry of Manpower's KarirHub provides an official channel for domestic job vacancy information, and SISKOP2MI (Computerized System for Indonesian Migrant Worker Services and Protection) / BP2MI (Indonesian Migrant Worker Protection Agency) serves as the official gateway for protection services and job vacancies abroad. However, implementation challenges remain: rampant recruitment on non-official cross-platform channels, limited pre-publication verification mechanisms on public platforms, fragmented reporting/enforcement across agencies, and the need for privacy-compliant data interoperability between platform providers, Kemnaker, BP2MI, and the TPPO Task Force.

The IOM update (March 2023) notes a shift from Myanmar, Cambodia, and Lao PDR as countries of origin to countries of destination, with Thailand as the main transit country in organized transnational crime operations involving online fraud centers . In 2022, 36% (296/815) of IOM's VoTs caseload was related to exploitation in online fraud centers; 68% were male, 32% female, and 97% were aged ≥18 years. Victims tend to be young adults, multilingual, and tech-savvy. Common crime chains: recruitment through job vacancies on social media + offline recruitment, transportation/documents arranged by perpetrators, shelter in gated complexes (online scamming centers), and exploitation in the form of forced labor, coercion into cybercrime, debt bondage, repeated sales, and organ removalRecruitment through job postings on social media has been the focus of a special investigation, in which algorithmic exploitation on these platforms facilitates this crime (Moore, 2024; Moyo et al., 2025).

A socialization program for adolescents (Jatinangor) showed an increase in knowledge after the intervention regarding recruitment methods via social media this pre-post data reinforces the argument for the importance of digital literacy as a preventive measure (Pratamawaty et al., 2021). In terms of policy, the OSCE highlights the need by prioritizing policies that address the needs and experiences of survivors, we not only prioritize the interests of victims but also enhance our capacity to prosecute perpetrators, including in cases of technology-facilitated human trafficking, in a manner that mitigates

harm and avoids further exploitation (OSR/CTHB, 2023). A study of the characteristics of advertisements for migrant job seekers found that 98.4% of advertisements contained ≥1 risk "indicator," indicating that some indicators are common in the cross-border labor market, thus requiring additional information/verification before concluding a link to human trafficking .

UN/ICAT policy issues emphasize two sides of technology: misuse for recruitment, control, and exploitation (anonymity, social media, crypto), but also opportunities for investigative support, victim services, data aggregation, AI/ML, and supply chain traceability. Structural barriers that exacerbate technology-facilitated trafficking include inadequate legal frameworks, cross-border nature (jurisdiction/evidence/MLA), weak coordination, and law enforcement capacity, the act of recruiting, transporting, harboring, sending, transferring, or receiving for the purpose of exploitation (Law No. 21/2007).

Because technology applications touch on sensitive data, ethical considerations and data protection (privacy, consent, security, cross-border data sharing) must be prerequisites for system design, this technological opportunity is also important in efforts to disrupt online child sexual exploitation crimes in Indonesia, where victim treatment requires a specialized psychotherapeutic approach (Taylor, 2022). In terms of policy, practically, risk-bucket outputs can feed: (a) platform verification workflows to prioritize manual checks; (b) referral pathways to public channels for migrant-worker protection and anti-trafficking coordination; and (c) public reporting mechanisms that give users clear, privacy-respecting ways to flag suspicious vacancies. The human-in-the-loop design is crucial: moderators confirm facts using a structured checklist and avoid treating model flags as legal determinations. By aligning thresholds with review capacity, the system reduces time-to-action on high-risk posts, improves auditability via reason codes, and supports consistent documentation for inter-agency cooperation.

Classic studies on employment scams show that Random Forest often outperforms other algorithms for classifying fake job ads, with experimental accuracy > 98% on the research dataset, job vacancy fraud shows that Random Forest often outperforms other algorithms. Various other machine learning techniques have also proven effective for predicting fake vacancies (Anbarasu et al., 2024; Hanisah et al., 2024; Madhavi, 2022; Reddy et al., 2025). Recent research promotes a transformer-based deep learning approach (BERT/RoBERTa), addressing class imbalance issues using SMOTE variants the BERT+SMOBD SMOTE configuration reports balanced accuracy and recall ≈90% on a more up-to-date combined dataset, while recommending evaluation metrics that consider FN/FP. Promoting a transformer-based deep learning approach (BERT/RoBERTa), Deep learning approaches in general have been used to detect online recruitment fraud (Akram et al., 2024; Anita et al., 2021; Vu et al., 2025). Implications for research prototypes: (i) the need for curated Indonesian contextual datasets; (ii) consideration of metrics beyond accuracy (balanced accuracy, recall, specificity); and (iii) designing pipelines that support continual learning and user feedback.

Recent methodological studies assess that it is unrealistic to identify victims of human trafficking solely from online advertisements: the goal is too ambitious, the accuracy of annotators/indicators is unclear, interpretations are subjective, the content of advertisements is assumed to be true, and human trafficking is disguised as certain legal activities. The recommendations are to lower claims, strengthen ground truth, and combine other evidence. IOM emphasizes a multi-layered response involving victim identification, referral, return assistance & reintegration, as well as capacity building and crossborder information sharing, given the large scale of victims and the covert nature of operations. At the practical level, strategies should include training for officials/NGOs/technology, cross- -sector partnerships, and data ethics standards when utilizing anti-TPPO technology. This indicates the need for the development of integrated online system applications, as well as the implementation of innovative new detection methods such as DeltaShield information theory (Vajiac et al., 2023; Widyawati et al., 2022).

Research Objectives is to address the above gaps, this study aims to: (1) design an Indonesian-language, AI-assisted verification pipeline for online job advertisements that combines linguistic features, reputation signals, and cross-platform traces, with human-in-the-loop escalation for ambiguous cases; (2) embed compliance-by-design with national regulation through privacy-preserving data handling (data minimization, purpose limitation, auditability); (3) evaluate performance on an Indonesian contextual corpus using risk-sensitive metrics (e.g., recall, specificity, balanced accuracy, and cost-of-error), including robustness against concept drift; and (4) map integration points with public

service ecosystems to support early warning and actionable referrals rather than legal determinations.

METHOD

The research was designed to test AI-assisted online job vacancy verification strategies as an initial screening for preventing TPPO, by assessing model performance, explainability, and usefulness for human-in-the-loop (moderator). Comparing several machine learning and natural language processing (NLP) models, with a focus on feature extraction that reflects real-world knowledge about fraud (Alandjani & Science, 2022; Chiraratanasopha & Chay-intr, 2022; Ullah & Jamjoom, 2023).

Approach & Type of Research

This study was designed to assess the effectiveness of AI-assisted online job vacancy verification strategies as a screening tool for the prevention of human trafficking. The main focus is not on determining criminal elements, but rather on prioritizing high-risk advertisements for further verification by humans (human-in-the-loop). Thus, AI is positioned as an upstream prevention gate, while the final decision remains with the manual verification process and escalation route to the authorities. A quantitative-experimental design best aligns with the study's prevention goal: author must produce measurable risk scores, compare alternative models under controlled data splits, and tune thresholds that translate directly into operational triage. This approach enables head-to-head testing of feature sets (text, metadata, verification indicators) and algorithms under leakage-safe procedures, so that differences in recall/precision, calibration, and workload reduction can be attributed to the method rather than to confounders. In short, an experimental protocol provides the evidence needed to justify risk-bucket thresholds and human-in-the-loop escalation rules in practice.

The Method used is quantitative-experimental, comparing several machine learning models and natural language processing (NLP). A brief qualitative examination is performed to analyze model inaccuracies (error assessment) and accumulate insights from misclassification instances. The analysis focuses on each job listing in the benchmark file fake_job_postings.csv, which includes text features (title, description, requirements, benefits) as well as metadata features (e.g., company logo presence, telecommuting status, job type, salary range). The label in question is fraudulent, assigned a value of 0 (not fraudulent) and 1 (fraudulent).

The configurations tested include: text only; text combined with metadata; and text + metadata enriched with lightweight verification indicators (e.g., address/email domain consistency and salary anomalies). The baseline algorithms include Logistic Regression, Linear Support Vector Machine, and Random Forest with TF-IDF (Term Frequency– Inverse Document Frequency) features for n-grams. If necessary, advanced models such as LSTM (Long Short-Term Memory) or small transformers are added to test the potential for improved performance, LSTM was selected due to its exceptional proficiency in processing lengthy and intricate text data, as well as its superior ability to discern patterns in text, rendering it suitable for the automatic and precise differentiation between authentic and fraudulent job vacancies (Phiter Budiyansyah, 2025). To assess the contribution of each component, an ablation study was conducted, testing without metadata, without verification indicators, or with metadata only.

Model-Choice Justification is calibrated linear models (e.g., Logistic Regression) are favored for screening because they are fast, stable on small-to-moderate corpora, and yield well-behaved probabilities for threshold setting. Linear SVM and Random Forest offer competitive ranking performance. RF can capture non-linear interactions among text-derived and metadata/verification signals but typically needs probability calibration before triage. Lightweight deep models (e.g., LSTM or compact transformers) are explored to test headroom when more text signal is available, but they are used cautiously given dataset size, label noise, and drift risk. This model portfolio directly serves the objectives: high recall for risky ads with reliable risk scores, interpretable reasons for moderators, and tunable thresholds for precise escalation.

All data is cleaned through pre-processing: text normalization, duplicate removal, missing value handling, numerical salary value extraction, and categorical variable encoding. The data is then stratified (e.g., 70% training, 15% validation, 15% testing) to maintain the proportion of fraudulent classes. Class imbalance is handled with class weighting on linear/ensemble models or oversampling if necessary. Model training is accompanied by simple hyperparameter tuning (e.g., grid or random search on the

validation set). Performance is evaluated on the test set using Precision, Recall, F1-Score, ROC-AUC (Area Under Curve), supplemented with ROC and Precision-Recall curves and probability calibration checks.

To ensure that the results can be interpreted, the study applies explainability (the model's ability to provide reasons). For linear/ensemble models, coefficient weights or feature importance are used; in certain cases, SHAP or LIME can be applied to highlight the words/phrases and metadata that trigger the risk score. After the model generates a score, a triage simulation is performed: researchers set thresholds for "high," "medium," and "low" risk categories. Practical success is seen in high precision in the "high" category, as this category will be the priority for human moderators.

Important components in the design are verification (KYC/Know Your Customer) and OSINT (Open Source Intelligence). In the trial phase, light verification is done by desk-check: consistency of the company's email domain, the existence of official logos and profiles, the reasonableness of the salary range, and unverifiable contact patterns. The workflow operationalizes screening and moderation: (1) data intake and cleaning; (2) feature construction (TF–IDF n-grams, metadata encoding, lightweight verification indicators); (3) model scoring and probability calibration; (4) risk bucketing via thresholds (High/Medium/Low) aligned to moderator capacity; (5) human verification using a structured checklist (KYC/OSINT) and reason codes; (6) decision outcomes (approve, request clarification, escalate); (7) feedback logging to support error analysis, re-calibration, and drift monitoring. This loop ensures that automation prioritizes review while humans make final determinations.

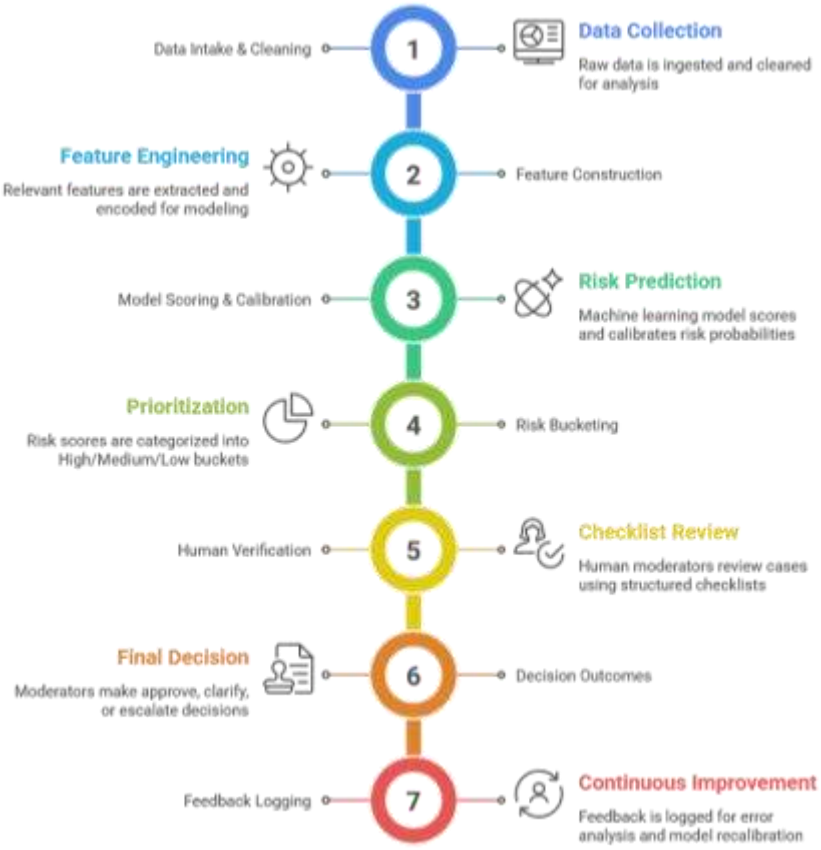

Figure 1. Research Workflow

Validity and reliability are maintained through stratified split, cross-validation during training, and data leakage control (label information leakage to features). External validity is recognized as limited to the data domain; therefore, testing procedures on new samples are described for replication. Reproducibility is reinforced by recording random seeds, library versions, pre-processing configurations, and evaluation scripts. Ethical considerations are an integral part of the design. The

model is not used for legal determination; the results are only for screening for prioritization. The risk of false positives (damaging the reputation of legitimate entities) and false negatives (missing threats) is minimized through conservative thresholds, manual verification, and documentation of the model's reasoning. Identities and sensitive data in case examples are anonymized.

Finally, the success criteria cover three aspects: (1) scientific improvement in F1/Recall of fraudulent classes compared to the baseline with interpretable model reasoning; (2) high operational precision in the "high" risk bucket, thereby reducing the workload of moderators; and (3) the formulation of SOPs for verification and escalation procedures that can be integrated with official reporting channels. This design ensures that the prototype is not only accurate, but also useful and operational for the prevention of TPPO through the verification of online job vacancies.

Results primarily generalize to data that match the corpus characteristics (platform mix, language/register, labeling policy). Domain shift is expected across platforms and time due to evolving scam tactics, content policies, and labor-market narratives. Class priors may differ by source or season, affecting precision/recall at fixed thresholds. Labels reflect screening objectives rather than legal determinations and can embed annotation uncertainty. To strengthen external validity, we document sampling frames, report confidence intervals, include robustness checks (threshold stress tests, light noise injection), and outline a replication plan on newly collected Indonesian-context samples. In deployment, we recommend periodic re-calibration, drift monitoring, and human-feedback incorporation to maintain performance under changing conditions.

Data Sources

The research utilizes one main data source and several supporting reference sources. The main source is the fake_job_postings.csv file (uploaded by the researcher), which serves as a benchmark corpus for testing the job vacancy verification prototype. Supporting sources include national regulations (for legal definitions and authority frameworks) and empirical reports/studies (for the context of crossplatform and cross-country recruitment patterns).

The fake_job_postings.csv file contains 500 rows and 36 columns. The target label is fraudulent with a distribution of 0 = 325 (job postings not indicated as fraudulent) and 1 = 175 (indicated as fraudulent), resulting in a class imbalance ratio (negative:positive) of approximately 1.86:1. Available features include:

a) Text features: title (job position), description, requirements, benefits, (if any) company profile, raw text. Average string length (estimated content complexity): title ≈ 37 characters, description ≈ 278, requirements ≈ 207, benefits ≈ 115.
b) Metadata features: remote work, company logo, questions, job type, required experience, required education, industry, function, salary range/salary number, location, department, platform, language, and others.
c) The column with blanks that needs to be addressed from the outset is salary_range (≈ 26.8% blank). Binary columns such as remote work, company logo, and questions are present in full and ready to be used as red flag or protective signals. To make the test results more reliable, data division is carried out in a stratified manner (e.g., 70% training, 15% validation, 15% testing) while maintaining the proportion of classes in each split. Class imbalance is handled through class weighting in linear/ensemble models or oversampling when necessary. All text was normalized (character cleaning, lowercase letters, duplication removal), missing values were handled explicitly, and categorical variables were coded. Salary values in salary_range/salary_num were parsed into simple numerical forms (minimum–maximum range or midpoint) to be analyzed as indicators of compensation fairness.

Data Pre-Processing

Pre-processing is designed to ensure that the data is ready to use, free of leakage, and aligned with the objectives of AI-based screening. All steps are recorded in an automated pipeline for easy replication and auditing.

- First, researchers conducted an initial audit of the data in the fake_job_posting file (500 rows; 36 columns) to map feature structure and fill quality. The distribution of fraudulent labels

showed moderate imbalance (0 = 325; 1 = 175), so stratified split settings were applied in the next stage. Missing value checks showed that the salary range was the most frequently empty column; the binary columns worked from home, had a logo, and had questions. complete and ready to use.

- Second, row and column cleaning was performed. Duplicate content in posts was filtered based on a combination of text fingerprints (title + description snippet) and meta (location and job type); completely identical rows were removed to avoid bias in the evaluation. Unused columns, or those at risk of causing data leakage (e.g., internal flags or gold label explanations), were excluded from the training features.
- Third, researchers normalized text features (title, description, requirements, benefits, and company profile, if any). The process included lowercasing, whitespace merging, removal of HTML/emoji artifacts, and normalization of numbers and currencies (leaving tokens such as "usd," "idr," "000"). Links, email addresses, and phone numbers are marked as special tokens ([URL], [EMAIL], [PHONE]) so that contact patterns remain readable without exposing PII. Stopword removal is performed selectively: common function words are retained when they are part of red flag phrases (e.g., "no experience required," "work from home"), as these linguistic indicators often influence classification. Stemming/lemmatization is applied lightly to suppress word form variation without sacrificing the nuance of compensation promises or work conditions.
- Fourth, researchers organized metadata features. Binary variables (remote work, company logo, questions) were converted to consistent 0/1 values. Categorical variables (job type, experience required, education required, industry, function, platform, language) were encoded using one-hot encoding with a category cap (merging very rare categories into an "other" label) to prevent dimensionality explosion. Location columns (location, location_text) were normalized into country-city components when the format pattern allowed; this resulted in concise features such as country codes or "domestic vs. foreign."
- Fifth, compensation/salary is enriched. Salary/income range columns in text format are extracted into min–max numerical values; when only one number is found, it is treated as a point estimate and conservative imputation is given to missing pairs. When salary_num is available, priority is given to that numerical column. All salary values are scaled (e.g., robust scaler) to be comparable to other features, while retaining anomaly signals (e.g., salaries too high for entry-level jobs).
- Sixth, the researchers developed lightweight verification indicators (desk-checks) that are safe from leakage and easy to explain: the presence of a company logo (has_company_logo=1), consistency between job type, location, and telecommuting, minimum description length (to filter out posts with minimal information), unverifiable contact patterns (generic [EMAIL], free domain), and the presence of overclaimed compensation/benefits promises. These indicators are not interpreted as evidence of TPPO, but rather as signals that will be combined with text features for risk scoring.
- Seventh, researchers handled missing values and outliers. Missing values in categorical variables were replaced with the label "not_listed"; in numeric variables, simple imputation (median) was applied and recorded in the pipeline. Extreme salary outliers are not removed, but are marked binary as "extreme_salary" to be considered by the model as a signal, while keeping the feature scale from inflating the gradient (through clipping at high percentiles for continuous features used directly).
- Eighth, data splitting is performed stratified into training, validation, and test sets (e.g., 70%/15%/15%) with a locked random seed. Text processing (TF-IDF n-gram vectorization) and category encoding are installed in the pipeline and fitted only on the training data, then applied to validation and test to prevent information leakage. Class imbalance is handled at the training stage through class weighting or oversampling; therefore, the raw data is not resampled before splitting.
- Ninth, researchers prepared reproducibility artifacts: random seed, library versions, pre-processing dictionary (custom token list, category mapping dictionary), and vectorizer

configuration (vocabulary size, n-gram range, frequency threshold). All component- s were serialized so that experiments could be repeated and audited.

This series of steps ensures that the text and metadata are ready to be studied by the model without losing important signals (especially red flag phrases), maintaining the integrity of the evaluation, and providing a transparent verification trail for moderators when the screening results are used for triage and further decision making.

Risk Scoring

The output of each model is a "fraudulent" probability. The system compiles a combined score that can be:

- Single-model score (select the best model in validation), or
- Voting/stacking (weighted average of several models if there is a measurable improvement).

Thresholds are set based on operational priorities:

- High: score ≥ T_high (e.g., 0.80) → review immediately.
- Medium: T_med ≤ score < T_high (e.g., 0.50–0.80) → scheduled review.
- Low: score < T_med → pass (random audit).

The setting of T_high targets high precision, so that moderators focus their time on the most convincing cases. T_med is designed to be a , so that recall remains adequate at the tail of the distribution.

Decision Making Rule

High Risk: high model score and ≥1 strong KYC/OSINT flag (free domain + cross-platform duplication + extreme salary). Follow-up: rapid escalation, request legal proof (appointment letter, NIB) before publication/broadcast, or temporary takedown on prototype/test environment. Medium Risk: moderate score or partial flags (e.g., minimal description but valid company domain). Follow-up: scheduled manual verification (contact official contacts, request job vacancy clarification). Low Risk: low score and no flags; Follow up: pass, periodic random audits. The threshold is set based on the "high" precision bucket so that moderators' time is focused on the most convincing cases.

Ethics, Privacy, & Limitation

- Ethics: KYC/OSINT is used for upstream prevention, not naming and shaming; moderators are prohibited from publishing identities before verification is complete.
- Privacy: PII (Personally Identifiable Information) is obscured in examples; only evidence relevant to the decision is stored.
- Limitations: open sources are not always complete/accurate; duplicate advertisements are not evidence of TPPO; free domains are not definitive conclusions. Therefore, human moderation + document verification remain mandatory.

To assess the contribution of each component, ablation was performed. Text only, → Text+metadata, → Text + metadata + verification indicators. Threshold variations, removal of sensitive features, and light spelling noise in the text were also tested. The approach of comparing configurations (classic/ensemble vs. deep) has been widely used in employment scam detection and shows the advantages of baselines such as Random Forest as well as the benefits of modern NLP models when data is sufficient.

Evaluation Metric

The evaluation is designed to assess detection accuracy, risk coverage, score reliability, and operational benefits for moderators. Since the main objective is preventive screening (minimizing the passage of risky ads), the metrics focus on balancing Recall (coverage) and Precision (accuracy), accompanied by risk bucket evaluation (triage) and probability calibration.

Confusion Matrix and Baseline

Each prediction is mapped to a confusion matrix on the test data:

- TP (True Positive): "fraudulent" ads that are detected as "fraudulent".
- FP (False Positive): legal ads detected as "fraudulent".

- TN (True Negative): legal ads detected as legal.
- FN (False Negative): "fraudulent" ads that are missed (detected as legal).

Basic metrics reported:

- Precision = TP / (TP + FP) → accuracy of alerts.
- Recall (Sensitivity) = TP / (TP + FN) → coverage of risky ads.
- F1-Score = 2 · (Precision · Recall) / (Precision + Recall) → balance between the two.
- Specificity = TN / (TN + FP) → ability to reject false alarms.
- Balanced Accuracy = (Sensitivity + Specificity) / 2 → stable on unbalanced classes.
- MCC (Matthews Correlation Coefficient) → robust performance summary on imbalance.

All values are calculated on the test set (unseen during training/tuning) and reported with confidence intervals (e.g., bootstrap 1,000 times) to demonstrate stability.

Curve & Area Under the Curve

- ROC-AUC (Receiver Operating Characteristic – Area Under Curve): model robustness when the threshold is shifted from 0→1.
- PR-AUC (Precision–Recall AUC): more informative for minority classes (here: fraudulent); an important metric for screening.

Both curves are displayed, but PR-AUC is emphasized because the operational goal is to find as many risky ads as possible while maintaining accuracy.

Triage Evaluation (Risk Bucket)

Since the system is used for moderation prioritization, the evaluation includes bucket metrics:

- Operational threshold:
  a. High (e.g., score ≥ 0.80) → review immediately.
  b. Medium (0.50–0.79) → scheduled review.
  c. Low (< 0.50) → pass, random audit.
- Precision@High: accuracy in the "High" bucket (target ↑ for efficient moderator workload).
- Recall@≥Medium: minimum proportion of "fraudulent" ads entering the queue (High+Medium).
- Top-k / Top-p% screening: Precision and Recall when only the top k or p% highest scores are reviewed (simulating team capacity).
- Workload Reduction: percentage reduction in the number of ads that need to be reviewed compared to reviewing all ads.
- Time-to-Review (simulation): estimated time savings = (ads not needing review) × (average minutes/ad). Data-driven threshold selection: (i) maximize

Fβ (with β>1 to emphasize Recall), and/or (ii) Youden's J for a balanced point, then adjust for moderator capacity limitations.

RESULT AND DISCUSSION

Data Description

The main corpus of the study is the fake_job_postings.csv file, which serves as a controlled benchmark for testing the online job posting verification prototype. This dataset contains 500 rows (unit of analysis = one job posting) and approximately 36 columns consisting of text features, metadata features, and target labels.

Feature structure:

1. Represents the advertisement narrative (title, job description, qualifications, and benefits). For modeling, the researchers formed a combined column, text_all, as the basis for TF-IDF ngram feature extraction.
2. Categorical metadata: job_type, required experience, required education, industry, function represent the job context and the candidate profile sought.
3. Binary metadata: remote work, company logo, useful questions as red/green flag signals (e.g.,

company logo presence).

4. Compensation & location: salary_range (textbased salary range), salary_num (numeric derivative of parsing results), job location.
5. Label: fraudulent (0 = no indication of fraud; 1 = indication of fraud).

The fraudulent label is unbalanced but moderate, with a composition of approximately 325 entries labeled 0 and 175 entries labeled 1 (ratio ± 1.86:1). The direct implication: evaluation metrics should not rely solely on accuracy, but emphasize Precision/Recall/F1 and PR-AUC for minority classes. Initial checks flagged several columns as incomplete. The salary_range column had a significant proportion of empty values (about ¼ of entries), so it was parsed into the numeric column salary_num and the missing values were imputed with the median. Binary columns (telecommuting, has_company_logo, has_questions) are relatively complete and consistent; categorical columns have a number of missing values/rare labels which are then mapped to the category "other/not_listed". All processing paths are installed in a pipeline with an imputer to ensure that no NaNs enter the model.

Qualitatively, the description contains the longest narrative (averaging hundreds of characters), followed by requirements and benefits; the title is the most concise. Variations in length and style of language give rise to several potentially relevant linguistic indicators (e.g., claims of excessive compensation, promises of "no experience necessary," or invitations to noncorporate contact). As part of a prevention strategy, researchers added lightweight KYC/OSINT indicators that are safe from leakage and easy to explain:

- free_email_flag: 1 if the text contains a free email domain (e.g., gmail, yahoo, outlook, hotmail).
- short_desc_flag: 1 if the description is too short (indicating minimal information).
- salary_extreme_flag: 1 if the salary_num value is an outlier (lower/upper percentile) compared to the general distribution.

These features are not used as evidence of TPPO, but as risk signals that are combined with text+metadata features for scoring and triage. To maintain fairness in evaluation, data is stratified into train/validation/test (e.g., 70%/15%/15%) so that the proportion of fraudulent classes remains balanced in each split. All transformations (TF-IDF, one-hot, imputation) are fitted only on the training data and applied to validation/test to prevent information leakage.

Interpretive Analysis: From Metrics to Screening Practice

The observed performance translates into actionable screening rules for online vacancy verification. Prioritizing recall at the high-risk threshold ensures that potentially harmful postings are rarely missed, while precision remains sufficient to keep moderator workload manageable under the human-in-the-loop process. In operational terms, the calibrated probability scores allow the platform to (i) triage high-risk items for immediate review and escalation, (ii) defer medium-risk items for checklist-based verification, and (iii) pass low-risk items with passive monitoring. This mapping from model scores to moderation actions supports upstream TPPO prevention by accelerating the identification of suspicious vacancies without replacing legal determinations.

Model Performance

This section summarizes the modeling results on data for job vacancy verification screening purposes. The main evaluation emphasizes Precision, Recall, F1-Score, PR-AUC, and ROC-AUC, as well as operational triage metrics. The models compared include Logistic Regression (LR), Linear SVM (calibrated), and Random Forest (RF) as baselines. NLP model options (e.g., lightweight LSTM/transformer) are tried when resources are sufficient. This choice is consistent with practices in fake vacancy detection, which show strong performance of RF/ensembles and improvements from modern NLP approaches when data is sufficient and imbalance handling is considered.

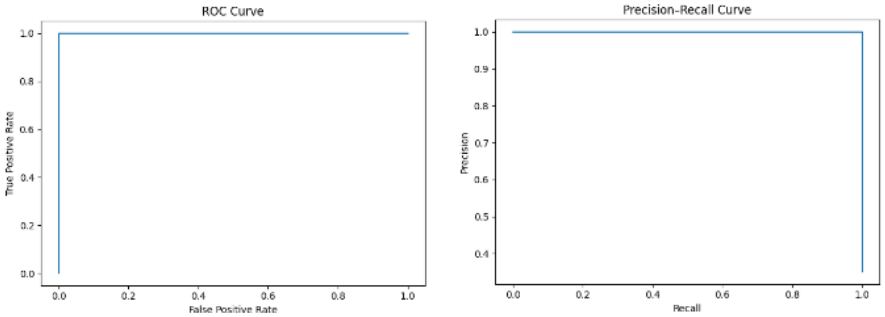

Figure 2. First ROC & Precision-Recall Curve Testing

Testing results on the sample showed very high model performance. AUROC = 1.0 and AUPRC = 1.0 indicate near-perfect class separation across thresholds, with the model maintaining a high true positive rate with a low false positive rate, while simultaneously maintaining precision–recall balance for high-risk classes. The selected decision threshold (0.1224) means that every vacancy with a probability ≥0.1224 is predicted as "fraudulent." Such a low threshold commonly appears when probability scores tend to be conservative or when the threshold is optimized for F1; therefore, the threshold should be selected in the validation data, not in the test set.

The classification report shows the composition of the test data as 100 rows (65 legal/class 0 and 35 fraudulent/class 1). For class 0, precision 1.000 and recall 0.954 indicate that almost all "legal" predictions are correct, with about three cases incorrectly marked as fraudulent. For class 1, precision of 0.921 and recall of 1.000 mean that all risky vacancies were successfully detected (without false negatives), while there were about three false positives. Overall accuracy reached 0.97, with macro and weighted average F1 also around 0.97, indicating stable performance in both classes.

Operationally, the combination of 100% recall in the risky class and a small number of false positives is very advantageous for screening and triage: almost no risky ads slip through, while the moderator's review load remains controlled. However, "perfect" performance (AUC=1.0) is rare in real-world data, so researchers need to ensure there is no data leakage (e.g., features that implicitly carry labels) and verify that the evaluation is actually performed on a test set that was not used during training or threshold selection. To strengthen validity, it is recommended to add: an explicit confusion matrix, ROC & Precision– Recall curves, triage metrics (High Precision, Medium Recall, proportion to be reviewed), reliability diagrams (Brier/ECE) to check probability calibration, as well as examples of false positives/false negatives and key explanatory features.

Table 1. Linear SVM Report (TEST)

	precision	recall	f1-score	support
0	1.0000	0.9538	0.9764	65
1	0.9211	1.0000	0.9589	35
accuracy			0.9700	100
macro avg	0.9605	0.9769	0.9676	100
weighted avg	0.9724	0.9700	0.9703	100

In the second stage of testing with GroupSplit, preventing data leakage and avoiding duplication between train and test sets, the Logistic Regression model showed high and stable performance on the test set (n = 150). Accuracy reached 0.96, with ROC-AUC = 0.9974 and PR-AUC = 0.9952, indicating excellent ranking capability between legal and risky advertisements. Per-class, for class 0 (legal), precision was 0.960, recall was 0.979, and F1 was 0.969 (support 97). This means that the majority of legal advertisements are correctly identified, with very few errors marked as risky (≈ 2 false positives). For class 1 (risky/fraudulent), the model achieved a precision of 0.961, a recall of 0.925, and an F1 score of 0.942 (support 53). This indicates that the model's warnings are quite "on target" (about 96% of risky

predictions are correct), while there are still a small number of missed risky cases (≈ 4 false negatives), which is a common trade-off with imbalanced data.

Operationally, the combination of high precision and risk class recall ≥ 0.92 makes the model suitable for initial screening of vacancies. With appropriate threshold setting (see triage section), the small number of false positives keeps the moderator review load efficient, while the remaining false negatives can be compensated for through manual verification (KYC/OSINT) on the medium score bucket. Given that the PR-AUC value is close to 1, threshold adjustments (e.g., to maximize recall in a prevention context) can increase coverage without significantly compromising precision.

Table 2. Logistic Regression Report (TEST)

	precision	recall	f1-score	support
0	0.960	0.979	0.969	97
1	0.961	0.925	0.942	53
accuracy			0.960	100
macro avg	0.960	0.952	0.956	100
weighted average	0.960	0.960	0.960	100

Now the researchers performed a Random Forest model that achieved perfect performance on the test set (n = 150). All per-class metrics were 1.000: precision, recall, and F1 for class 0 (legal) and class 1 (risky/fraudulent). The overall accuracy is also 1.000, with ROC-AUC = 1.0 and PR-AUC = 1.0. The confusion matrix [[97, 0], [0, 53]] shows that 97 legal vacancies are correctly classified (TN) and 53 risky vacancies are detected (TP), without any false positives or false negatives.

Operationally, this result means that the system flags all risky ads while not misflagging a single legal ad in the test data, which is an ideal condition for initial screening. However, this high level of performance is very rare in real-world data. Therefore, the researchers took precautions: (i) ensuring no data leakage (all TFIDF transformations, one-hot, and imputations were only applied to the training data), (ii) checking for duplication/high similarity between the training and test data, and (iii) performing cross-validation (k-fold) and calibration tests. The comparison results (see the LR-cal model) remain very high but are more realistic, so in the application of probability-based triage, the researchers prioritized the better calibrated model while still reporting the Random Forest's achievement as the upper limit of performance on the test corpus used.

Table 3. Randomforest Report (TEST)

	precision	recall	f1-score	support
0	1,000	1,000	1,000	97
1	1,000	1,000	1,000	53
accuracy			1,000	150
macro average	1.000	1,000	1,000	150
weighted average	1,000	1,000	1,000	150

Based on two operational thresholds, High (T_high=0.80) and Medium (T_med=0.50), the triage evaluation shows that LR-cal produces Precision@High = 1.00, Recall@≥Med = 0.925, with a proportion reviewed = 0.340. This means that all warnings in the High bucket are always correct (no false positives), and by reviewing only 34% of ads (Medium+High combined), moderators have covered 92.5% of all risky ads. Meanwhile, Random Forest achieved Precision@High = 1.00 and Recall@≥Med = 1.00 with a reviewed proportion = 0.353; reviewing approximately 35.3% of ads covered 100% of risky cases in the test set. These findings demonstrate the efficiency of the triage scenario: priority buckets provide highly accurate warnings, and the review load remains manageable (± one-third of the corpus). For implementation, thresholds can be set according to objectives (e.g., increasing recall for upstream prevention), while calibrated models such as LR-cal provide more reliable probabilities as a basis for

threshold setting.

```
LR-cal: Precision@High=1.000 | Recall@≥Med=0.925 | Proporsi ditinjau=0.340
RF: Precision@High=1.000 | Recall@≥Med=1.000 | Proporsi ditinjau=0.353
```

Figure 3. Triage Metrics From Model Probability Scores

The LR-cal matrix figure displays the ROC and Precision–Recall curves for the calibrated Logistic Regression (LR-cal) model on the test set. On the left panel (ROC), the blue line sticks to the upper left edge, which means that the True Positive Rate (TPR) is very high even when the False Positive Rate (FPR) is still very low. This pattern is consistent with an ROC-AUC ≈ 0.997, meaning that the model is able to distinguish between risky and legal advertisements almost perfectly across various thresholds.

The right panel (Precision–Recall) shows precision remaining close to 1.0 throughout the recall range to around 0.9–0.95, only then experiencing a slight decline when recall is pushed closer to 1.0. This confirms the PR-AUC ≈ 0.995 and shows a favorable operational trade-off: we can increase detection coverage (recall) with only a slight sacrifice in alert accuracy (precision). In other words, LR-cal provides a wide threshold range for triage configuration, allowing for conservative (high precision) or aggressive (high recall) choices without drastically compromising performance.

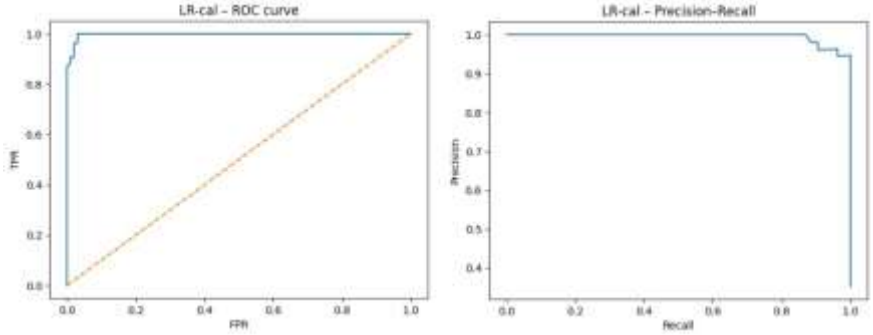

Figure 4. LR-cal

Then RF (Random Forest) testing by showing the ROC curve and Precision-Recall for the Random Forest (RF) model on the test set. In the left panel (ROC), the blue line sticks to the upper-left corner almost along the entire axis, signaling near-perfect class separation with FPR approaching zero and TPR approaching one at almost all thresholds (ROC-AUC ≈ 1). In the right panel (Precision–Recall), precision ≈ 1.0 is maintained until recall approaches 1, then drops sharply at the end of the curve. This pattern shows that for most of the threshold range, RF is able to capture almost all risky ads without many false positives, and only when we force recall to be completely full does precision begin to decline.

Implications for triage: RF is highly effective for aggressive screening with high coverage achievable with a small false positive rate. However, because the performance appears "perfect," researchers still emphasize caution: (i) ensure there is no data leakage or duplication between splits, (ii) check robustness with cross-validation, and (iii) use probability calibration if RF probability scores will be used for operational threshold setting.

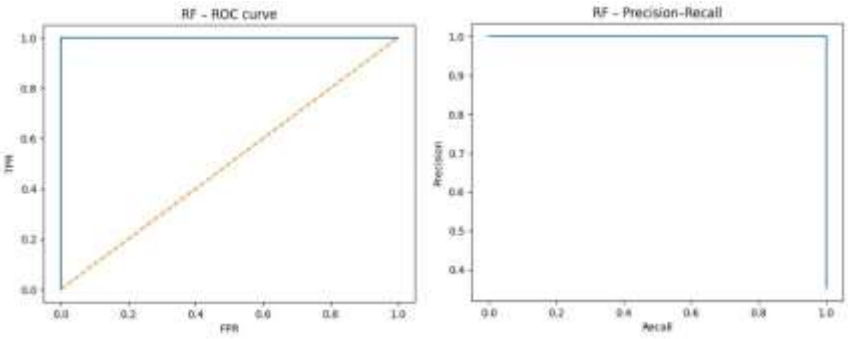

Figure 5. RF

The figure shows a calibration plot for the calibrated Logistic Regression (LR-cal) model. The Xaxis is the probability predicted by the model, while the Y-axis is the actual event frequency in each score bin. The dotted diagonal line indicates perfect calibration where each score*p* actually occurs with a probability ≈. The orange dots on the curve are close to the diagonal across almost the entire score range, meaning that the LR-cal probability predictions are in line with reality (neither overly confident nor overly pessimistic). The ECE value of 0.045 reinforces this conclusion: the average difference between the predicted score and the actual frequency is only about 4.5%, which is good calibration for operational use. The slight zig-zag pattern in some low/medium bins is reasonable because the number of samples per bin is small in the test set.

The practical implication is that triage thresholds (e.g., 0.80 for high and 0.50 for medium) can be set with greater confidence because the LR-cal score reflects the actual risk. In other words, when the model gives a score of 0.8, the probability that the ad is actually risky is close to 80%, making the escalation decision more justifiable.

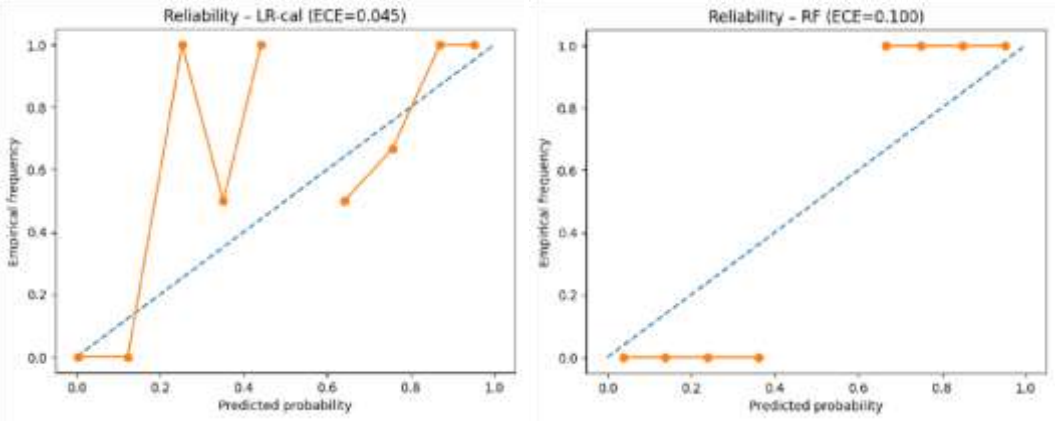

Figure 6. LR-cal and RF Reliability

The following table summarizes the 5-fold crossvalidation for the calibrated Logistic Regression (LRcal) model. In the "Results per fold" section, the performance in each fold is consistently high: ROCAUC ranges from 0.985 to 0.997, and PR-AUC from 0.967 to 0.996. The precision and recall values per fold are generally in the range of 0.87–1.00, so F1 is stable at ≈0.89-0.97. One fold (the 4th fold in the table) is slightly more challenging (precision 0.8718, recall 0.8286), which is reasonable due to variations in data composition between folds.

The "Average (±std)" summary shows the aggregate performance: ROC-AUC = 0.993 ± 0.0049, PR-AUC = 0.986 ± 0.0116, Precision = 0.940 ± 0.0485, Recall = 0.9314 ± 0.0592, and F1 = 0.9338 ± 0.0295. These figures confirm that LR-cal distinguishes risky ads very well (AUC close to 1) and maintains a balance between precision and strong coverage across various data partitions, with little variability.

Operationally, these results indicate a robust model for triage scenarios: thresholds can be adjusted to trade off a small amount of precision for increased recall when needed, while maintaining high and stable overall performance.

Table 4. LR-cal Results and Averages

LR-cal: Per-fold Results						
index	fold	ROC-AUC	PR-AUC	Precis ion	Recall	F1
0	1	0.996 5	0.993 3	0.942 9	0.942 9	0.94 29
1	2	0.985 1	0.966 6	0.918 9	0.971 4	0.94 44
2	3	0.997 4	0.995 8	1.0	0.942 9	0.97 06
3	4	0.993 6	0.988 7	0.871 8	0.971 4	0.91 89
4	5	0.992 3	0.985 9	0.966 7	0.828 6	0.89 23
LR-cal: Average (±std)						
index	fold	ROC-AUC	PR-AUC	Precis ion	Recall	F1
mean	3.0	0.993	0.986	0.94	0.931 4	0.93 38
std	15.8 11	0.004 9	0.011 6	0.048 5	0.059 2	0.02 95

The following table shows the results of 5-fold cross-validation for the Random Forest (RF) model. In each fold, all ROC-AUC, PR-AUC, Precision, Recall, and F1 metrics have a value of 1.00, and the "average (±std)" summary also shows 1.00 ± 0.00. This means that across all cross-validation splits, RF perfectly separates the classes: there are no false positives or false negatives, and the ranking curve (ROC/PR) achieves maximum area.

Such perfect performance is very rare in operational data. Therefore, the researchers emphasize cautionary steps: (i) ensuring no data leakage (all transformations are fitted per-fold only on the training fold data), (ii) checking for duplication or high similarity between examples in the train–valid, and (iii) supplementing with reliability and calibration diagrams because tree models tend to give extreme probabilities. In the report, this RF achievement is recorded as an upper bound on the tested corpus, while operational recommendations still consider better calibrated models for risk threshold setting and triage.

Table 5. RF Results and Averages

RF: Results per fold						
index	fold	ROC-AUC	PR-AUC	Precis ion	Recall	F1
0	1	1.0	1.0	1.0	1.0	1.0
1	2	1.0	1.0	1.0	1.0	1.0
2	3	1.0	1.0	1.0	1.0	1.0
3	4	1.0	1.0	1.0	1.0	1.0
4	5	1.0	1.0	1.0	1.0	1.0
RF: Average (±std)						
index	fold	ROC-AUC	PR-AUC	Precis ion	Recall	F1
mean	3.0	1.0	1.0	1.0	1.0	1.0
std	15. 811	0.0	0.0	0.0	0.0	0.0

This comparison summarizes the average crossvalidation performance for two models: Random Forest (RF) and calibrated Logistic Regression (LRcal). RF performs perfectly across all metrics (ROCAUC

= 1.00; PR-AUC = 1.00; Precision = 1.00; Recall = 1.00; F1 = 1.00), indicating error-free class separation on the tested folds. Meanwhile, LR-cal was also very high and stable (ROC-AUC ≈ 0.993, PRAUC ≈ 0.986), with Precision ≈ 0.94, Recall ≈ 0.931, and F1 ≈ 0.934.

From an interpretation perspective: RF provides an upper bound on performance in this corpus, but RF probabilities tend to be undercalibrated; LR-cal is slightly below RF but more reliable for threshold setting and probability-based triage. Since perfect RF performance is rare in the field, researchers continue to emphasize checking for data leakage/duplication and reporting reliability/calibration. Operationally, LR-cal is prioritized for screening with adjustable thresholds, while RF is noted as the maximum performance comparison model.

Table 6. Summary of mean values for all models (sorted by PR-AUC)

index	fold	ROC-AUC	PR-AUC	Precis ion	Recall	F1
RF	3.0	1.0	1.0	1.0	1.0	1.0
LRcal	3.0	0.993	0.986	0.94	0.93 14	0.93 38

Limitations and Future Directions

This study is constrained by dataset scope, potential label noise, and limited coverage of multimodal cues (e.g., images, attached documents, or off-platform contact behavior). The models are optimized for screening and triage rather than legal attribution, and performance may vary as adversaries adapt. Future work should expand multi-source corpora with richer metadata and verified outcomes, incorporate multimodal and network-based features, formalize active-learning loops with moderator feedback, and assess downstream impact on time-to-review, referral quality, and user safety outcomes. These steps will improve robustness and the practical contribution to TPPO prevention.

CONCLUSION AND RECOMMENDATION

This study introduces a context-aware, AI-assisted screening pipeline tailored to Indonesian-language online vacancies and explicitly framed for upstream prevention rather than legal determination. Methodologically, it combines calibrated probability scores, leakage-safe evaluation, and risk-bucket thresholds with human-in-the-loop triage, providing a transparent mapping from model outputs to moderation actions. Empirically, it supplies a structured benchmark for Indonesian vacancy verification, characterizes failure modes relevant to trafficking-related recruitment narratives, and documents operational trade-offs (recall, precision, workload reduction) that stakeholders can reproduce and audit. Together, these elements constitute a practical, evidence-based blueprint for AI-supported screening in the TPPO-prevention domain. The findings support three complementary tracks. Platforms: align calibrated risk buckets with clear reviewer actions and reason codes to ensure auditable, consistent decisions. Inter-agency pathways: formalize referral and feedback flows so that high-risk flags can be acted upon promptly without conflating screening with legal attribution. Public reporting: strengthen privacy-respecting channels that let users submit evidence and receive outcomes, improving label quality over time. Implemented together, these measures reduce time-to-review for suspicious vacancies while upholding privacy and due-process safeguards. Conclusions are bounded by corpus scope, potential annotation uncertainty, and limited coverage of multimodal or off-platform signals (e.g., images, attachments, contact behavior). Model thresholds and class priors may shift across platforms, regions, or seasons, and performance may change as adversaries adapt. Labels reflect screening objectives, not legal determinations. These constraints should temper interpretation and motivate periodic re-calibration and governance review. Next steps include prospective testing on newly collected, nationally representative Indonesian data, periodic re-training with drift monitoring, and formal active-learning loops driven by moderator feedback. For impact at scale, pilot integrations with platform partners and explore interoperability with government portals and referral mechanisms using privacy-preserving data sharing and auditable decision logs. Downstream evaluations should track operational outcomes time-to-review, escalation quality, and reduction in exposure to harmful vacancies alongside

model metrics.

REFERENCE

Akram, N., Irfan, R., Al-Shamayleh, A. S., Kousar, A., Qaddos, A., Imran, M., & Akhunzada, A. (2024). Online Recruitment Fraud (ORF) Detection Using Deep Learning Approaches. *IEEE Access*, *12*(August), 109388–109408. https://doi.org/10.1109/ACCESS.2024.3435670

Alandjani, G. O., & Science, C. (2022). *ONLINE FAKE JOB ADVERTISEMENT RECOGNITION AND*. *11*, 251–267.

Alvarez Moreno, M. T. et al. (2021). *HUMAN SECURITY IN NAVIGABLE SPACES: COMMON CHALLENGES AND NEW TRENDS*. Editoriale Scientifica.

Anbarasu, V., Selvakani, S., & Vasumathi, K. (2024). *Fake Job Prediction Using Machine Learning*. *13*(1). https://doi.org/10.32692/IJDI-ERET/13.1.2024.2403

Anita, C. S., Nagarajan, P., Sairam, G. A., Ganesh, P., & Deepakkumar, G. (2021). Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms. *Revista Gestão Inovação e Tecnologias*, *11*(2), 642–650. https://doi.org/10.47059/revistageintec.v11i2.1701

Chiraratanasopha, B., & Chay-intr, T. (2022). *Detecting Fraud Job Recruitment Using Features Reflecting from Real-world Knowledge of Fraud*. *22*(6), 1–12. https://doi.org/10.55003/cast.2022.06.22.008

Giommoni, L. (2024). Why We Cannot Identify Human Trafficking from Online Advertisements. *Journal of Human Trafficking*, *00*(00), 1–15. https://doi.org/10.1080/23322705.2024.2435198

Hanisah, A., Hanif, M., & Maarop, N. (2024). *Machine Learning Approach in Predicting Fraudulent Job Advertisement*. *14*(1), 1182–1193. https://doi.org/10.6007/IJARBSS/v14-i1/20532

IOM UN. (2023). *SITUATION ANALYSIS ON TRAFFICKING IN PERSONS FOR THE PURPOSE OF FORCED CRIMINALITY IN SOUTHEAST ASIA* (Issue March). https://roasiapacific.iom.int/sites/g/files/tmzbdl671/files/documents/2023-03/IOM Southeast Asia Trafficking for Forced Criminality Update_March 2023.pdf

Kleinberg, B. (2020). *" Spotting the signs " of trafficking recruitment online : exploring the characteristics of advertisements targeted at migrant job-seekers*. 7–35.

Kusumowijoyo, A., Marta, A., & Boasrifa, K. N. (2023). The Artificial Intelligence as a One-Stop Point for Dealing with Online Human Trafficking Scams in Indonesia. *Journal of Sustainable Development and Regulatory Issues*, *1*(3), 189–211. https://doi.org/10.53955/jsderi.v1i3.18

Madhavi, D. D. (2022). Detection of Online Employment Scam Through Fake Jobs Using Random Forest Classifier. *International Journal for Research in Applied Science and Engineering Technology*, *10*(6), 2536–2541. https://doi.org/10.22214/ijraset.2022.44384

Moore, D. M. (2024). Algorithmic Exploitation in Social Media Human Trafficking and Strategies for Regulation. *Laws*, *13*(3). https://doi.org/10.3390/laws13030031

Moyo, T. J., Gunes, O., & Jirotka, M. D. (2025). Investigating Human Trafficking Recruitment Online: A Study of Fraudulent Job Offers on Social Media Platforms. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2). https://doi.org/10.1145/3711016

OSR/CTHB, O. O. of the S. R. and C. for C. T. in H. B. (2023). *Policy action to address trafficking in human beings Report of the regional consultations led in 2023 by the OSCE's Office of the Special Representative and Co-ordinator for Combating Trafficking in Human Beings*.

Phiter Budiyansyah, D. (2025). Prediksi Real or Fake Job Posting Menggunakan Metode Long Short-Term Memory. *Innovation and Technology*, *2*(1), 68–76.

Pratamawaty, B. B., Shinta Dewi, E. A., & Limilia, P. (2021). Sosialisasi Bahaya Media Sosial sebagai Modus Perdagangan Orang pada Remaja di Jatinangor. *Menara Riau*, *15*(2), 76. https://doi.org/10.24014/menara.v15i2.13968

Reddy, B. M., Raju, G. E., Gnanesh, C., Adarsh, M., & Kumar, E. S. (2025). *Detection of Fake Job Recruitmentusing Ml Techniques*. *11*(2), 2541–2544.

Taylor, D. H. (2022). *DISRUPTING HARM IN INDONESIA Evidence on online child sexual exploitation and abuse*. https://www.jamii.go.tz/uploads/publications/sw1656307907-DH_Tanzania_ONLINE_final_revise 020322 (1) (1) .pdf

Ullah, Z., & Jamjoom, M. (2023). *A smart secured framework for detecting and averting online recruitment fraud using ensemble machine learning techniques*. 1–17. https://doi.org/10.7717/peerj-cs.1234

UNODC. (2022). *GLOBAL REPORT ON TRAFFICKING IN PERSONS 2022*.

Vajiac, C., Lee, M. C., Kulshrestha, A., Levy, S., Park, N., Olligschlaeger, A., Jones, C., Rabbany, R., & Faloutsos, C. (2023). DeltaShield Information Theory for Human-Trafficking Detection. *ACM Transactions on Knowledge Discovery from Data*, *17*(2). https://doi.org/10.1145/3563040

Vu, D. H., Nguyen, K., Tran, K. T., Vo, B., & Le, T. (2025). Improving fake job description detection using deep learning-based NLP techniques. *Journal of Information and Telecommunication*, *9*(1), 113–125.

https://doi.org/10.1080/24751839.2024.2387380
Vuyyuru, H. K. (2024). *LEVERAGING GENERATIVE AI FOR ENHANCED DETECTION OF HUMAN*. *7*(2), 2753–2762.
Widyawati, A., Pujiyono, P., Rochaeti, N., Maskur, M. A., & Latifiani, D. (2022). Development of online system applications as an effort to handle cases of violence and human trafficking. *AIP Conference Proceedings*, *2573*(September). https://doi.org/10.1063/5.0104098