



Benchmarking Supervised Learning, Contextual Transformers, and Explainability Layers for Online Recruitment Fraud Detection

Prof. Sanjana Pawale, Venkatesh Gajul, Vaishnavi Nigde, Sourik Porya

ASM's Institute of Business Management and Research, Pune, Maharashtra, India

Corresponding Author Email: sanjanapawale@asmedu.org, venkateshgajul783@gmail.com,
vaishnavinigade04@gmail.com, sourikporya@gmail.com

How to Cite this Article:

Gajul, V., Nigde, V. & Porya, S. (2026). Benchmarking Supervised Learning, Contextual Transformers, and Explainability Layers for Online Recruitment Fraud Detection. International Journal of Creative and Open Research in Engineering and Management, 2(10), 1-9. <https://doi.org/10.55041/ijcope.v2i10.027>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i10.027>

Abstract—

Digital hiring platforms make applicant sourcing fast and scalable, but they also give bad actors an easy way to distribute fraudulent job postings. Detecting Online Recruitment Fraud (ORF) automatically remains difficult because scam texts mimic authentic corporate ads, and natural datasets suffer from severe class skew (scams make up less than 5% of postings in the wild). In this study, we survey recent detection techniques and benchmark several model families—from traditional shallow baselines and recurrent nets to bidirectional Transformers—against the 17,883-record Employment Scam Aegean Dataset (EMSCAD). We evaluate Word2Vec semantic projections, address minority-class starvation with SMOTE oversampling, and assess model interpretability via SHAP and LIME. Regularized XGBoost paired with SMOTE and dense Word2Vec features yielded 99.44% accuracy, a 0.99 F1-score, and an AUC of 0.9999. Pretrained Transformer encoders (RoBERTa and BERT) also performed strongly (98.81% and 98.67% accuracy, respectively) while better tolerating linguistic variations. Because deep neural nets and boosted tree ensembles do not provide transparent reasoning out of the box, we couple these models with local attribution plots to build an auditable inspection pipeline for real-world hiring platforms.

Keywords— Online Recruitment Fraud, Transformer Architectures, BERT, XGBoost, Explainable AI, SMOTE, EMSCAD, Hiring Security.



I. INTRODUCTION

Modern recruitment runs almost entirely on cloud-hosted boards and professional social networks. While this shift has lowered recruitment overhead and broadened the applicant pool, it has also expanded the surface for automated fraud.

Online Recruitment Fraud (ORF) involves publishing fictitious vacancies designed to trick applicants into paying advance fees, buying dummy equipment, or disclosing sensitive identity records. These scams cause more than monetary loss. Human-factors research suggests cognitive biases can prevent applicants from recognizing warning signs in professional-looking listings. At the same time, when scammers clone legitimate company names, the victim company faces brand dilution and legal scrutiny. Consequently, dependable, low-latency screening tools are an important operational requirement.

Automated detection systems face three concrete engineering bottlenecks. First, linguistic mimicry allows fraudulent posts to copy corporate phrasing, Equal Opportunity Employer statements, and plausible job requirements, making basic keyword matching ineffective. Second, severe class skew causes unbalanced models to favor legitimate posts and miss fraudulent listings. Third, opaque decision paths make it difficult for security analysts to verify why boosted trees and deep Transformer encoders flagged a particular listing.

This paper reviews existing approaches across machine learning, recurrent architectures, and Transformers, then evaluates these models side-by-side using dense vectorization, SMOTE balancing, and SHAP/LIME explanation layers.

II. LITERATURE REVIEW

2.1 Classical Classifiers and Tabular Metadata

Early attempts to catch job scams relied on metadata verification and token heuristics. Prior work examined recruiter metadata mismatches and combined non-text variables—such as salary details, remote-work indicators, and logo uploads—with linguistic tokens to train Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) baselines. Feature choice strongly influences these traditional models. TF-IDF has been reported to outperform standard Bag-of-Words by reducing the weight of frequent stop words, while Chi-Square ranking and Principal Component Analysis (PCA) can reduce collinear noise and training cost. However, linear decision boundaries struggle with dense non-linear interactions; the reviewed SVM baseline achieved only about 50.44% accuracy and a 0.35 F1-score.

2.2 Deep Sequence Models and Balancing Approaches

To capture sentence structure, subsequent studies used recurrent neural networks. Bidirectional LSTMs reached 98% accuracy on EMSCAD, while CNN-Bi-LSTM and CNN-Bi-GRU hybrids were used to learn both local n-gram patterns and longer document context. Self-attention mechanisms were also applied to LSTM models to prioritize suspicious phrases such as wire-transfer and processing-fee language. Class imbalance remains central: Focal Loss can reduce the dominance of authentic postings during optimization, while ADASYN and SMOTE oversampling can improve minority recall.

2.3 Contextual Transformers and Explainability (XAI)

Modern fraud detection increasingly relies on pretrained Transformer architectures. Unlike static bag-of-words or directional recurrent networks, multi-head attention dynamically models token relationships according to sentence context. Fraud-BERT demonstrated the value of contextual pretraining for deceptive-language detection. Because deep neural networks hide their decision logic, interpretability tools have become important for production deployment. SHAP and LIME can generate token-attribution maps that explain why a specific posting was flagged, making automated detection more auditable.

III. METHODOLOGY

3.1 Dataset Insights and Empirical Heuristics

Experiments were conducted on the Employment Scam Aegean Dataset (EMSCAD), containing 17,883 job postings described by 18 attributes spanning text fields and categorical flags. The ground-truth distribution is severely skewed: 17,014 postings (95.14%) are legitimate, while 866 (4.86%) are fraudulent, producing an imbalance ratio of approximately 19:1.

Missing Company Logos: The `has_company_logo` flag shows a -0.26 correlation with fraudulent posts. Legitimate listings include verified logos in roughly 80% of cases, while scam postings frequently leave this field blank.

Omitted Compensation: The `salary_range` feature is disproportionately empty or listed as “0-0” among fraudulent listings, indicating that scammers often avoid verifiable compensation figures.

Experience Skew: Postings cluster heavily within “Entry level” and “Mid-Senior level” categories, indicating that fraudulent campaigns target both recent graduates and experienced professionals.



3.2 Preprocessing and Dense Vectorization

The title, company_profile, description, and requirements fields were merged into a single text block for each posting. Null values were imputed and duplicate listings were removed to prevent test-set leakage. Characters were converted to lowercase, HTML artifacts, punctuation, and extraneous spacing were stripped, and uninformative English stop words were removed. The cleaned documents were projected into dense continuous vectors using Word2Vec with an embedding size of $d = 300$, context window $c = 5$, minimum token frequency of 1, and four parallel worker threads.

3.3 Handling Class Skewness with SMOTE

SMOTE was applied strictly to the training fold to prevent decision boundaries from favoring the majority class. For each minority instance, SMOTE identifies its k -nearest minority neighbors in Euclidean space, selects a neighbor, and synthesizes an intermediate vector. This expands the minority decision boundary without simply duplicating identical samples.

3.4 Classification Architectures

Extreme Gradient Boosting (XGBoost): An ensemble of shallow decision trees trained iteratively with a second-order regularized objective to reduce overfitting on dense embedding vectors.

Contextual Transformers (BERT & RoBERTa): Pretrained Transformer models process input sequences bidirectionally using multi-head self-attention. BERT uses masked-token and next-sentence objectives, whereas RoBERTa removes the next-sentence objective, uses dynamic masking, and trains with larger batches.

3.5 Explainability Layer: SHAP and LIME

SHAP (Shapley Additive Explanations): Uses cooperative game theory to compute the marginal contribution of each feature across possible feature subsets. TreeExplainer was used with XGBoost for feature-importance analysis.

LIME (Local Interpretable Model-agnostic Explanations): Perturbs individual input texts, weights samples by proximity, and fits an interpretable local surrogate model to highlight suspicious indicators in a flagged post.

IV. RESULTS AND DISCUSSION

4.1 Evaluation Setup and Metrics

Experiments were run in Python 3.8 across the 17,883 EMSCAD records. Model quality was

measured using Accuracy, Precision, Recall, F1-Score, and AUC derived from the confusion matrix.

4.2 Impact of Train-Test Partitions

The 80:20 split produced a minor performance lift over the 70:30 split when models were paired with SMOTE. For XGBoost, accuracy was 99.44% with an 80:20 split versus 99.41% with a 70:30 split. RoBERTa achieved 98.81% versus 98.66%, respectively. The 80:20 partition was therefore used for subsequent benchmarks.

4.3 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1	AUC
XGBoost	99.44	0.99	0.99	0.99	0.9999
Random	99.37	0.99	0.99	0.99	—
RoBERTa	98.81	0.99	0.99	0.99	0.9991
BERT	98.67	0.99	0.99	0.99	—
Decision	96.97	0.97	0.97	0.97	—
LSTM	93.14	0.93	0.93	0.93	—
SVM	50.44	0.69	0.50	0.35	—

The XGBoost confusion matrix contained 3,365 true negatives, 31 false positives, 7 false negatives, and 3,403 true positives. The low false-negative count demonstrates high sensitivity to fraudulent postings. RoBERTa recorded 18 false negatives, compared with 7 for XGBoost, but its self-attention mechanism better handles linguistic variation and phrasing shifts. The RBF-kernel SVM failed to separate classes effectively in the dense vector space and produced near-random baseline performance.

4.4 Comparison with Prior Work

Earlier EMSCAD benchmarks reported approximately 95.40% accuracy for ensemble methods such as ORFDetector and 98.71% for Bi-LSTM implementations. The integration of XGBoost with Word2Vec and SMOTE achieved 99.44% accuracy, exceeding these baselines.

4.5 Practical Deployment Considerations

SMOTE Prevents False Comfort: Models trained on raw, unadjusted data can show deceptively high accuracy while performing poorly on the minority class. SMOTE brought XGBoost and RoBERTa precision and recall to 0.99.

Inference Latency vs. Semantic Depth: XGBoost offers low training overhead and millisecond-scale inference, making it suitable for high-volume batch processing.



RoBERTa captures subtle context shifts and is useful for real-time screening of unusual phrasing.

Auditable Screening via XAI: Pairing high-performing models with SHAP and LIME generates attribution maps that identify suspicious terms, such as unconventional wire-payment instructions, for administrative review.

V. CONCLUSION

Online Recruitment Fraud continues to exploit digital employment platforms, putting job seekers at financial and personal risk. This study evaluated classical classifiers, deep sequence models, and bidirectional Transformers on EMSCAD under severe class imbalance. Dense Word2Vec vectors paired with SMOTE oversampling and regularized XGBoost yielded 99.44% accuracy, a 0.99 F1-score, and an AUC of 0.9999. Pretrained Transformer encoders such as RoBERTa achieved 98.81% accuracy and a 0.9991 AUC, showing strong performance on unstructured text. Integrating SHAP and LIME improves interpretability and supports practical deployment where analysts need auditable explanations.

ACKNOWLEDGMENT

The authors acknowledge ASM's Institute of Business Management and Research, Pune, Maharashtra, India, for supporting this research.

REFERENCES

- [1] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, "Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset," *Future Internet*, vol. 9, no. 1, p. 6, 2017.
- [2] S. Lal, R. Jaiswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, "ORFDetector: Ensemble learning based online recruitment fraud detection," in *Proc. 12th Int. Conf. Contemporary Computing (IC3)*, 2019, pp. 1–5.
- [3] J. Lee and M. J. Cho, "Online job scams: Unveiling the impact of overconfidence, digital literacy, and algorithmic literacy on user susceptibility to false job advertisements," *New Media & Society*, 2025.
- [4] V. Anbarasu, S. Selvakani, and M. K. Vasumathi, "Fake job prediction using machine learning," *Ubiquity*, vol. 13, no. 1, pp. 12–20, 2024.
- [5] R. Rofik, R. A. Hakim, J. Unjung, B. Prasetyo, and M. A. Muslim, "Optimization of SVM and gradient boosting models using GridSearchCV in detecting fake job postings," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 419–430, 2024.
- [6] A. S. Pillai, "Detecting fake job postings using bidirectional LSTM," *Int. Res. J. Modern Engineering and Technology Science*, vol. 5, no. 3, pp. 1825–1830, 2023.
- [7] S. Chavhan, R. C. Dharmik, and S. Jain, "Evaluation of CNN-BiGRU and CNN-BiLSTM model for fake job post detection: A deep learning approach," in *Proc. 2nd Int. Conf. Emerging Trends in Engineering and Medical Sciences (ICETEMS)*, 2024.
- [8] S. Badere et al., "An intelligent system for identifying fake job ads using CNN-BiGRU and CNN-BiLSTM," in *Proc. 4th Int. Conf. Advancement in Electronics & Communication Engineering (AECE)*, 2024.
- [9] S. S. Sanisetty, S. V. Kotamaraja, B. N. Reddy, and S. Vekkot, "Comprehensive approach to fraudulent job post detection using machine learning and BERT models," in *Proc. 4th Int. Conf. Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 2025.
- [10] K. Taneja, J. Vashishtha, and S. Ratnoo, "Fraud-BERT: Transformer based context aware online recruitment fraud detection," *Discover Computing*, vol. 28, no. 1, p. 9, 2025.
- [11] C. Srikanth, M. Rashmi, S. Ramu, and R. M. Guddeti, "A novel fake job posting detection: An empirical study and performance evaluation using ML and ensemble techniques," in *Proc. Int. Conf. Security, Privacy and Data Analytics*, 2022.
- [12] F. G. Hussain et al., "Fake news detection landscape: Datasets, data modalities, AI approaches, their challenges, and future perspectives," *IEEE Access*, 2025.
- [13] A. Papasavva et al., "Applications of AI-based models for online fraud detection and analysis," *Crime Science*, vol. 14, no. 1, p. 7, 2025.
- [14] S. Safdar and M. Wasim, "DFN-SCNC: Detecting fake news based on social context and news content: A hybrid approach using BERT and Bi-GRU," in *Proc. Int. Conf. Frontiers of Information Technology (FIT)*, 2024.
- [15] M. Wasim et al., "KeepUp: A unified framework fusing knowledge extraction, social platform engagement, and user profiling for fake news detection," *Array*, vol. 29, p. 100687, 2026.
- [16] K. Yu, S. Jiao, and Z. Ma, "Fake news detection based on BERT multi-domain and multi-modal fusion network," *Computer Vision and Image Understanding*, vol. 252, p. 104301, 2025.



- [17] L. Caruccio et al., “Identifying fake reviews for refund purposes: Evaluating the effectiveness of a transfer-learning model against emerging large language models,” *Engineering Applications of Artificial Intelligence*, vol. 162, p. 112448, 2025.
- [18] R. Gupta, I. Kashyap, and V. Jindal, “SBiLM: Siamese Bi-LSTM model for handling imbalance in fake review detection,” *Procedia Computer Science*, vol. 235, pp. 1157–1166, 2024.
- [19] K. R. Reddy, G. Indrani, N. P. Kumar, and K. V. Krishna, “Fake job posting detection using machine learning algorithms,” in *Proc. 4th Int. Conf. Innovative Mechanisms for Industry Applications (ICIMIA)*, 2025.
- [20] A. J. Veliyath et al., “Fake job detection using statistical and NLP based analysis,” in *Proc. IEEE 15th Int. Conf. Electronics Information and Emergency Communication (ICEIEC)*, 2025.