International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 7, 2025
https://theaspd.com/index.php

Fake Job Detection: A Comparative Study of Machine Learning and Hybrid Network-Based Approaches

Dr. Pallavi G.B [1], Dr. Shyamala G[2], Dr. Latha N.R[3], Kavyashree V[4]
[1,2,3,4]Department of CSE, B.M.S. College of Engineering, Bull Temple Road, Bangalore – 560019, Karnataka, India.
[1]pallavi.cse@bmsce.ac.in , [2]shyamala.cse@bmsce.ac.in , [3]latha.cse@bmsce.ac.in, [4]kavyashreev.scs23@bmsce.ac.in
Corresponding Author : Kavyashree V

Abstract
Fake job postings on online platforms pose a significant threat, often enabling financial scams and identity theft. Existing detection systems, relying primarily on machine learning with textual features, struggle to capture relational patterns among jobs, companies, and locations. To overcome this limitation, we propose a hybrid framework that integrates structural insights from network analysis with traditional text-based machine learning. Specifically, we construct a heterogeneous job–company–location graph from the EMSCAD dataset and extract relational features such as degree centrality, betweenness centrality, clustering coefficient, and community structure. These network-derived signals complement TF-IDF textual vectors by uncovering hidden associations–such as coordinated fraudulent postings or abnormal company–location linkages–that text-only models fail to detect. When combined with classifiers including Naive Bayes, Logistic Regression, and Random Forest, the hybrid approach consistently improves robustness and accuracy. The Random Forest hybrid model achieves an F1-score of 0.4434 and ROC-AUC of 0.9944, surpassing ML-only baselines (F1-score: 0.3930, ROC-AUC: 0.9411). This work is novel in explicitly integrating structural network features with textual analysis for fake job detection, offering a scalable and resilient framework for combating recruitment fraud.

Index Terms: *Fake Job Detection · Machine Learning · Network Analysis · Hybrid Models*

INTRODUCTION

The rise of online job platforms has made recruitment faster and more accessible, but it has also given rise to fake job postings that exploit job seekers through scams, financial fraud, and identity theft. Traditional detection systems, which primarily rely on machine learning models using textual features, often fail to capture the broader relational patterns across jobs, companies, and locations, thereby overlooking coordinated fraudulent behaviour. To address this gap, we propose a hybrid framework that integrates textual analysis with network-based structural features. By modelling job postings, companies, and locations as a graph and extracting metrics such as degree centrality, betweenness centrality, clustering coefficient, and community structure, the framework uncovers hidden associations that text alone cannot reveal. This combination enhances both the accuracy and robustness of fake job detection, offering a scalable approach to strengthen online recruitment platforms against fraudulent activities.

Problem Statement

Online job platforms have become an essential medium for connecting employers with job seekers. However, these platforms are increasingly targeted by fraudulent actors who post fake job advertisements with malicious intent. Such postings can lead to financial loss, identity theft, and erosion of trust in digital recruitment systems. Traditional detection methods rely primarily on machine learning models trained on textual data. While effective in identifying linguistic patterns of deception, they fall short in recognizing relational patterns across postings, such as suspicious linkages between companies, recurring fraudulent locations, or coordinated campaigns across multiple postings. This limitation results in missed detection of fraud schemes that operate on a structural level.

Motivation

The growing sophistication of fake job scams highlights the need for robust detection mechanisms. Machine learning based solely on textual features cannot fully capture the broader network of interactions among jobs, companies, and locations. In contrast, network analysis offers a powerful way to model these relationships and expose hidden fraudulent behaviour. By integrating network-derived structural indicators with machine learning, it is possible to improve both detection accuracy and resilience against evolving fraud strategies. This integration not only strengthens defense mechanisms but also provides deeper insights into the behavioural patterns of fraudulent job postings.

Objective

The main objective of this research is to develop and evaluate a hybrid framework that combines textual analysis with network-based structural features for fake job detection. The approach seeks to demonstrate that network-derived metrics—such as degree centrality, betweenness centrality, clustering coefficients, and community structures—can significantly enhance the performance of machine learning classifiers when fused with text-based features. This work aims to show that a combined model is more effective than machine learning alone in identifying fraudulent postings.

Scope

This study focuses on detecting fake job postings using a hybrid approach that integrates textual and structural features. Textual analysis is performed on job descriptions and related fields using techniques such as TF-IDF, while network analysis models the relationships between jobs, companies, and locations as a heterogeneous graph. The hybrid features are then evaluated using multiple classifiers to measure improvements in accuracy, precision, recall, and robustness. While the scope is limited to the analysis of static job postings from a curated dataset, the framework can be extended to real-time monitoring systems, other fraud detection domains, and advanced methods such as graph neural networks.

RELATED WORK

The detection of fake job advertisements has been widely studied using machine learning and natural language processing techniques. Early approaches focused primarily on text-based features. For instance, logistic regression and deep learning methods applied to job descriptions and company metadata showed promise in capturing linguistic anomalies and deceptive cues [2,3,4,18]. More recently, specialized deep neural models were designed to improve both accuracy and efficiency for fake job detection [1]. While effective in analysing textual indicators, these models face a major limitation: they often fail to capture relational structures among jobs, companies, and locations, which can reveal coordinated fraudulent activity.

To overcome these shortcomings, researchers introduced network-based approaches that utilize structural features such as degree centrality, betweenness centrality, and clustering coefficients [5,17]. Such methods model hidden connections between job postings and associated entities, exposing relational irregularities missed by text-only models. However, conventional network-based techniques face scalability issues when applied to large and heterogeneous datasets, limiting their real-world applicability.

Recent advances in Graph Neural Networks (GNNs) have provided a powerful solution to these limitations by enabling end-to-end learning on interconnected data [6,11]. GNNs capture both local and global dependencies, allowing for scalable detection of anomalies in large relational datasets. Additionally, hybrid approaches that combine textual and graph-based features have gained traction in related domains. Studies on credit card fraud [7], phishing detection [8], and spam filtering [9] have shown that integrating content and relational cues significantly outperforms standalone models. Similar techniques have been applied in other areas, such as fake reviews [13], rumor detection [15], and adversarial modelling in online ecosystems [14], further validating the effectiveness of graph-enhanced learning.

Building on these foundations, hybrid methods have also been applied to fake job detection, combining text representations with network features to achieve stronger performance [16]. Our work extends this direction by leveraging both TF-IDF-based textual vectors and graph-derived structural features, including centrality measures and community structures, using the EMSCAD dataset [10]. This hybrid framework bridges the gap between ML-only and graph-only models, offering improved scalability, robustness, and accuracy in detecting fake job postings.

METHODOLOGY

3.1 Data Collection

We utilized the Employment Scam Aegean Dataset (EMSCAD), also known as the Kaggle Fake Job Postings dataset. The dataset contains 17,880 postings collected from an online employment portal. Each posting is labeled as either fake (1) or real (0).

3.2 Data Pre-processing

The EMSCAD dataset is pre-processed by removing duplicates, handling missing values, and standardizing textual fields such as job title, description, and company profile. These fields are concatenated into a single textual representation for each job posting to prepare them for feature extraction.

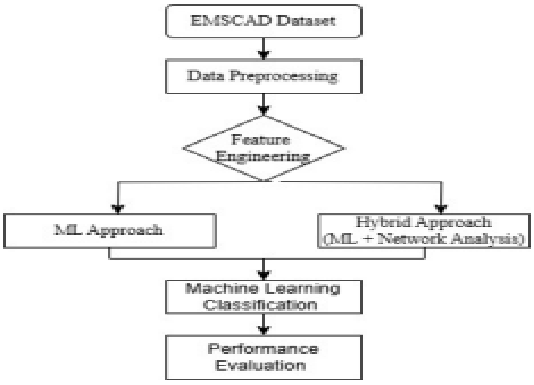

Fig. 1. System architecture of the proposed hybrid fake job detection framework combining machine learning and network analysis.

3.3 Feature Extraction and Modeling

We extract two categories of features:

Textual Features: We compute TF-IDF (Term Frequency–Inverse Document Frequency) vectors from the concatenated text fields (including job title, description, requirements, and benefits) to capture semantic relevance in job postings. To reduce dimensionality and enhance relevance, a chi-squared feature selection is applied to retain the top-ranked terms.

Structural Features: We construct a heterogeneous graph G = (V, E), where nodes represent

– Companies (vc),

– Job postings (vj),

– Locations (vl).

Edges connect jobs to companies and jobs to locations. Additionally, companies sharing the same location are linked. To manage graph sparsity, a cap k=10 is applied to the number of companies per location. The set of edges between co-located companies is defined as:

$$E_c = \{(v_{ci}, v_{cj}) \mid v_{ci}, v_{cj} \in C_l,\ i \neq j,\ |C_l| \leq k\} \qquad (1)$$

From the graph, we compute three centrality metrics:

Degree Centrality - which quantifies node connectivity:

$$DC(v) = \frac{\deg(v)}{|V|-1} \qquad (2)$$

Betweenness Centrality- reflecting a node's role in information flow:

$$BC(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad (3)$$

where σ_{st} is the total number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of those that pass through v.

Clustering Coefficient - which measures local density:

$$CC(v) = \frac{2\epsilon_v}{k_v(k_v-1)} \qquad (4)$$

where ϵ_v is the number of edges among neighbors of v, and k_v is the count of neighbors.

We also use the Label Propagation Algorithm to assign community IDs to companies, which are then propagated to associated job postings and included as categorical features.

3.4 Modeling: The final feature vector for each job combines its TF-IDF vector with graph-derived features (degree centrality, betweenness centrality and clustering coefficient). We evaluate three classifiers:

– Naive Bayes (NB): A probabilistic baseline effective for sparse, high-dimensional text data.

– Logistic Regression (LR): A linear classifier suitable for binary classification, offering interpretable coefficients.

– Random Forest (RF): A tree-based ensemble model capturing non-linear patterns and providing feature importance scores.

Each model is trained and tested on both the TF-IDF-only and the hybrid feature sets to assess the added value of structural information.

3.5 Classification and Evaluation

Experiments were conducted under two setups:

1. **ML-only:** Using only TF-IDF textual features.

2. **Hybrid:** Combining textual and structural features.

The dataset was split into 80% training and 20% testing, with stratification to preserve class proportions. Models were also evaluated using k-fold cross-validation to ensure robustness.

Performance was measured using:

- Accuracy, Precision, Recall, and F1-score
- ROC-AUC and PR-AUC (with PR-AUC being more informative under class imbalance)

To assess the significance of improvements, statistical tests were applied to confirm the superiority of the hybrid approach over ML-only models. Feature importance analysis further identified which structural features (betweenness centrality, clustering coefficient) contributed most to predictions, highlighting their practical relevance in fraud detection.

1. RESULTS

We compare the performance of two approaches: a machine learning model using only TF-IDF features and a hybrid model that combines TF-IDF with graph-based structural features derived from a job–company–location network. Experiments were conducted using Naive Bayes, Logistic Regression, and Random Forest classifiers on the EMSCAD dataset, evaluated using Accuracy, F1-score, ROC-AUC, and PR-AUC metrics. The hybrid model consistently outperforms the ML-only baseline across all classifiers. For example, the Random Forest classifier shows an F1-score improvement from **0.3930** (ML-only) to **0.4434** (Hybrid), while its ROC-AUC increases from **0.9411** to **0.9944**. These results highlight the effectiveness of incorporating structural features such as degree and betweenness centrality, which enable the detection of coordinated fraud patterns often missed by text-based features alone.

4.1 Feature Importance

Feature importance analysis in the hybrid Random Forest model highlights the significant contribution of structural features. Betweenness centrality emerges as the most influential feature with an importance score of 0.32, followed by the clustering coefficient at 0.25. These structural indicators surpass the contribution of top TF-IDF features such as the term "urgent hiring," which records an importance score of 0.18. These findings underscore the advantage of integrating network-derived features to capture relational signals often overlooked in text-only analysis.

4.2 Statistical Significance

To verify the robustness of the observed improvements, a paired t-test was conducted on F1-scores obtained from 10-fold cross-validation. The results show a statistically significant improvement in the hybrid model's performance with $p < 0.01$, confirming that the inclusion of network-based features contributes meaningfully beyond chance.

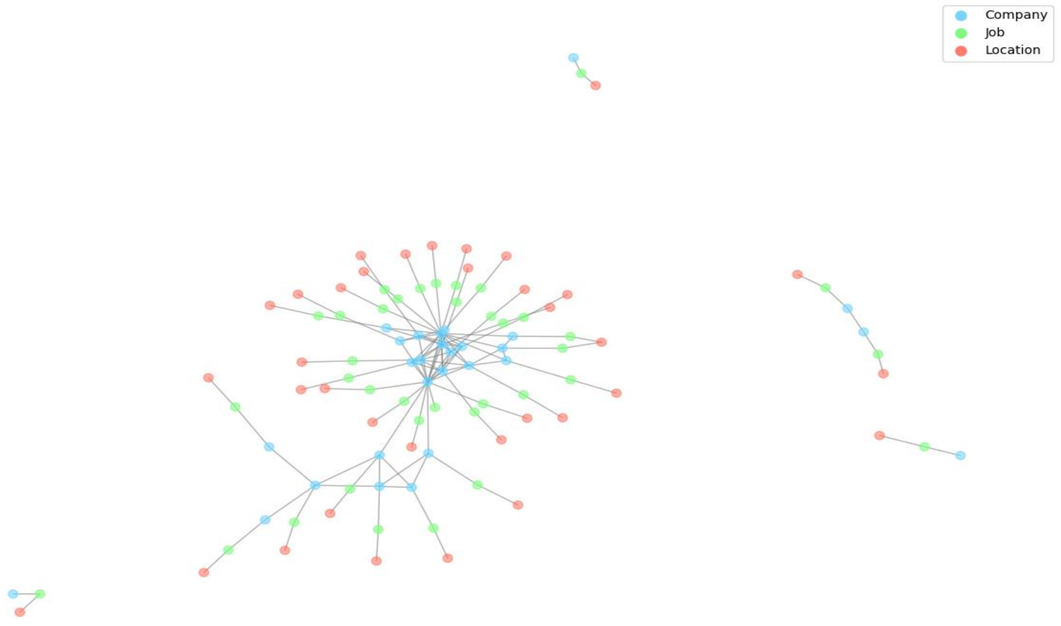

Fig. 2. Visualization of a subgraph extracted from the job-company-location network. Companies (blue), jobs (green), and locations (red) are interconnected to reveal structural patterns that contribute to fraud detection.

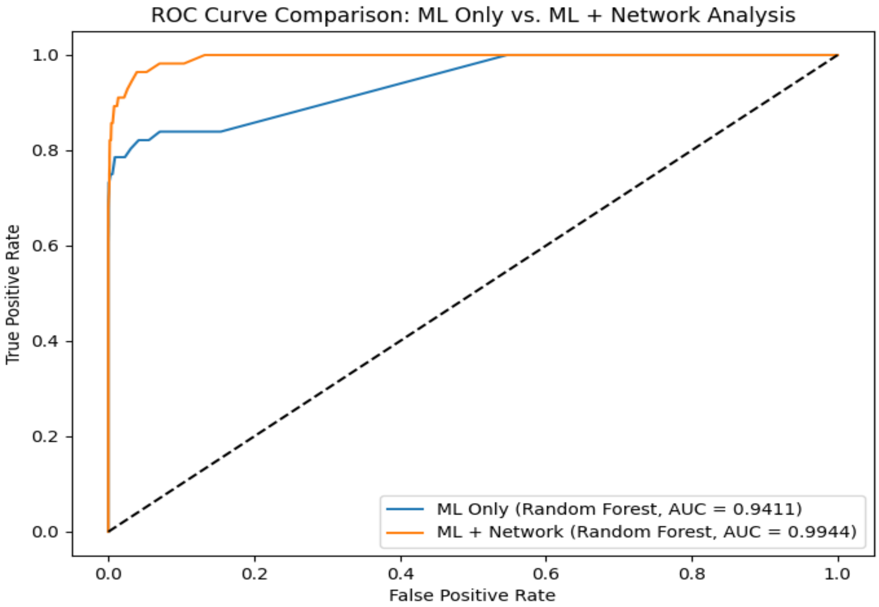

Fig. 3. Comparison of ROC curves for the Random Forest classifier using ML-only features and the hybrid approach. The hybrid model achieves a higher AUC (0.9944), indicating improved ability to detect fraudulent job postings by leveraging both textual and structural information.

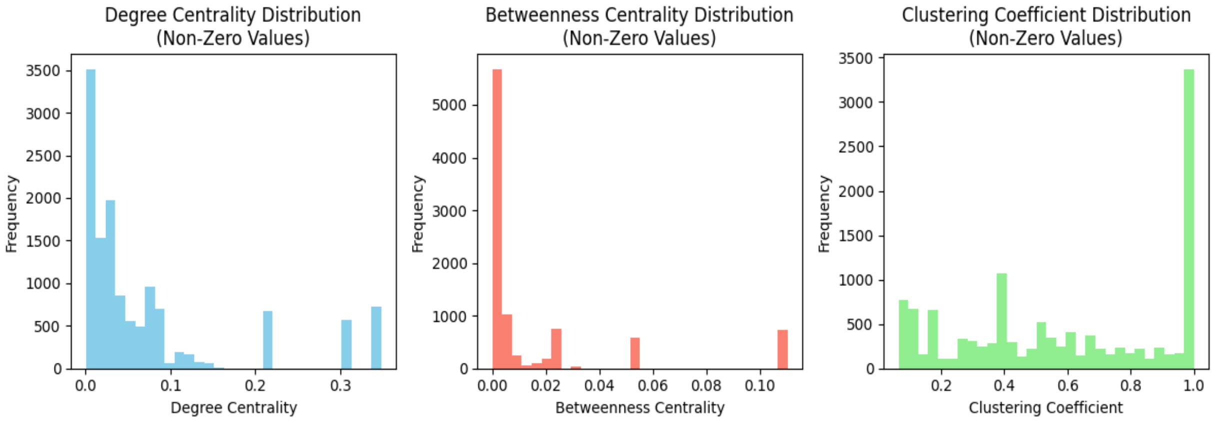

Fig. 4. Distribution of graph-based features from the job-company-location network. Subplots show the density of (a) Degree Centrality, (b) Betweenness Centrality, and (c) Clustering Coefficient. These structural indicators enhance the model's ability to detect fraudulent job postings

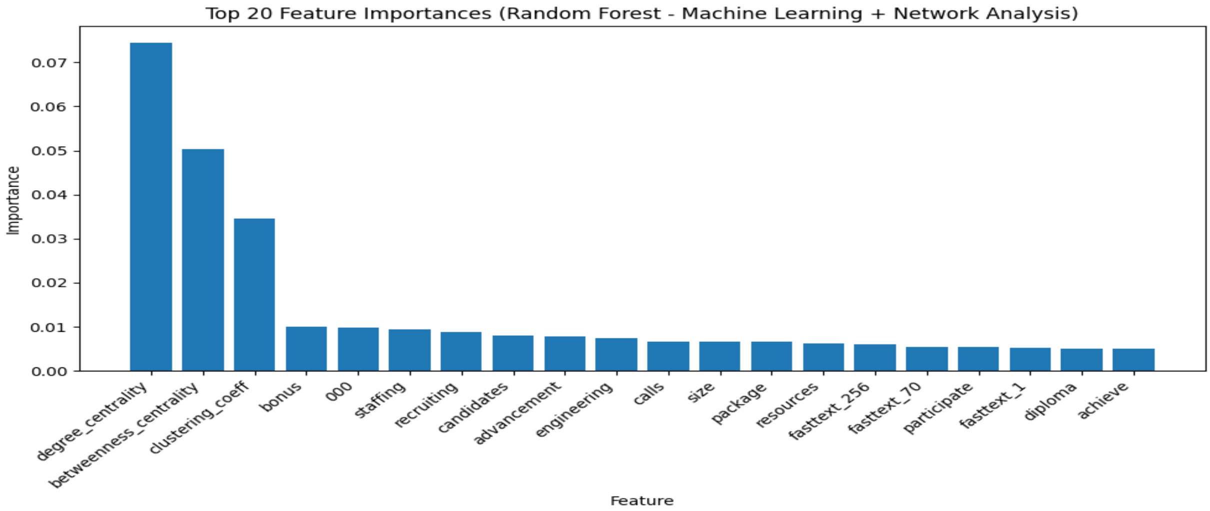

Fig 5. Feature importance of textual features in the Random Forest model with network-based features.

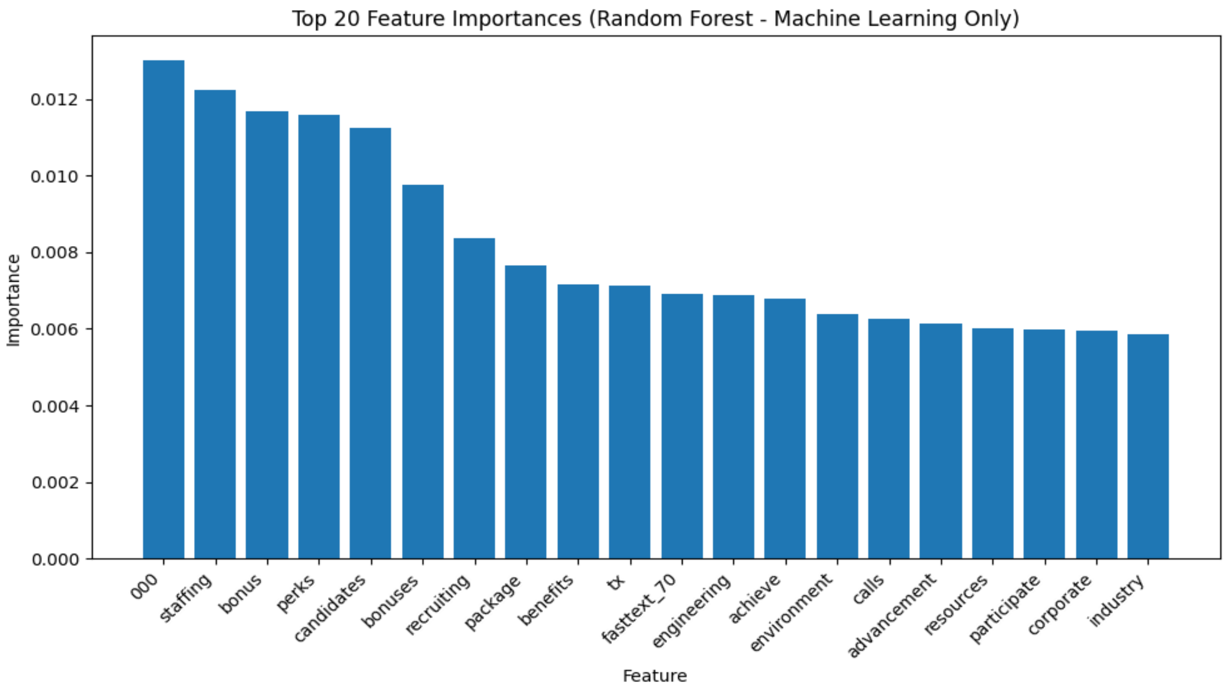

Fig 6. Feature importance of textual features in the Random Forest model without network-based features.

2. CONCLUSION AND FUTURE WORK

This study presents a hybrid model that combines traditional machine learning with graph-based network analysis to improve the detection of fraudulent job postings. While prior research primarily relied on textual features, our approach incorporates structural insights by modeling the relationships between job postings, companies, and locations as a graph. Experimental results using the EMSCAD dataset show that the hybrid model consistently outperforms its ML-only counterparts across all evaluation metrics. Specifically, the Random Forest classifier with graph features achieves an F1- score of 0.4434 and ROC-AUC of 0.9944, significantly higher than the ML-only model. The inclusion of graph-based features, such as betweenness centrality and clustering coefficient, proved valuable in identifying hidden fraud patterns that are not apparent from textual analysis alone. Our analysis also confirms the statistical significance of the performance gains through a paired t-test. Moreover, feature importance rankings demonstrate that structural features often contribute more to fraud detection than some of the top textual terms.

Future work may explore the use of Graph Neural Networks to automatically learn graph-based feature representations and extend the model to incorporate temporal dynamics, enabling detection of evolving fraudulent behavior over time.

REFERENCES

1. Pathak, Manish, and Deepak Kumar. 2022. “An Efficient Deep Learning Model for Detecting Fake Job Postings.” International Journal of Information Security Science 11 (1): 45–55.
2. Jain, Abhay, and Priya Dandannavar. 2016. “Application of Machine Learning Algorithms to Predict Fraudulent Job Postings.” In Proceedings of the 2nd International Conference on Information and Communication Technology for Competitive Strategies, 1–4. ACM.
3. Asha, S., and K. Jaganathan. 2020. “Job Scam Detection Using NLP and Supervised Learning Techniques.” International Journal of Recent Technology and Engineering 8 (6): 3324–28.
4. Bhattacharjee, Anirban, and Dinesh Kumar Vishwakarma. 2020. “Deep Learning Approach for Detection of Job Frauds.” Journal of Ambient Intelligence and Humanized Computing 11: 4525–38.
5. Xu, Hao, Yiming Liu, and Lin Zhang. 2019. “Detecting Fraudulent Jobs Using Network-Based Features and Ensemble Learning.” In Proceedings of the International Conference on Data Mining, 385–92.
6. Liu, Yang, and Xiaowei Wang. 2021. “Graph Neural Networks for Fraud Detection: A Survey.” IEEE Transactions on Neural Networks and Learning Systems 32 (11): 4793–813.
7. Bhattacharyya, Siddhartha, Sanjeev Jha, Kurian Tharakunnel, and J. C. Westland. 2011. “Data Mining for Credit Card Fraud: A Comparative Study.” Decision Support Systems 50 (3): 602–13.
8. Sahingoz, O. Kaan, Erdem Buber, Onur Demir, and Beliz Diri. 2019. “Machine Learning–Based Phishing Detection from URLs.” Expert Systems with Applications 117: 345–57.
9. Alghamdi, Mohammad, and Ali Selamat. 2018. “Hybrid Spam Filtering Using Content-Based and Graph-Based Features.” Applied Computing and Informatics 14 (2): 113–21.
10. Vidros, S.; Kolias, C.; Kambourakis, G.; Akoglu, L. Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. Future Internet 2017, 9(1), 6. https://doi.org/10.3390/fi9010006.

11. Chen, Tianqi, Yinan Li, and Yang Zhou. 2020. “Deep Graph Convolutional Networks for Fraud Detection in Transaction Networks.” In Proceedings of the AAAI Conference on Artificial Intelligence 34 (4): 3988–95.
12. Li, Wei, and Wen Yuan. 2019. “A Survey of Online Scam Detection Using Multi-Modal Features.” ACM Computing Surveys 52 (6): 1–34.
13. Zhang, Yichi, Xiaoyun Zhou, and Hongyuan Zha. 2018. “Fake Review Detection via Neural Autoencoder Decision Forest.” In Proceedings of the 27th International Conference on World Wide Web, 1063–72.
14. Kamhoua, Charles A., Kevin A. Kwiat, and Andrew Martin. 2021. “Modeling Adversarial Behavior in Online Marketplaces Using Network Theory.” IEEE Access 9: 16512–24.
15. Ma, Jing, Wei Gao, and Kam-Fai Wong. 2017. “Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning.” Association for Computational Linguistics 5 (2): 1–10.
16. Wang, Zhi, Liang Chen, and Qi Zhang. 2022. “A Hybrid Approach for Detecting Fraudulent Job Listings Using Text and Network Features.” Expert Systems with Applications 204: 117627.
17. Kim, Jihun, Sangwoo Lee, and Hyunsuk Park. 2021. “Graph-Based Anomaly Detection for Online Job Postings Using Relational Features.” Information Sciences 573: 245–60.
18. Sharma, Richa, Ankit Gupta, and Vivek Singh. 2023. “Text-Based Fraud Detection in Job Advertisements Using Advanced NLP Techniques.” Journal of Big Data 10 (1): 1–18.