IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

Smart Detection of Fraudulent Online Job Posting Using Machine Learning

Ms. Yasmin Rahim Shaikh[1] and Prof. Ms. Swati S. Hinge[2]
[1]PG Student, Sanghavi College of Engineering, Varvandi, Nashik, Maharashtra
[2]Assistant Professor, Sanghavi College of Engineering, Varvandi, Nashik, Maharashtra

Abstract: *The rapid growth of online recruitment platforms has significantly simplified the hiring process but has also led to a sharp rise in fraudulent job postings. Such deceptive advertisements exploit job seekers by extracting sensitive information or financial payments under false pretenses. Manual verification of job postings is impractical at scale, necessitating automated detection mechanisms. This paper proposes a machine learning–based framework for the intelligent detection of fraudulent online job advertisements. The Employment Scam Aegean Dataset (EMSCAD) is used as the benchmark dataset. Natural Language Processing (NLP) techniques with Term Frequency–Inverse Document Frequency (TF-IDF) are employed for textual feature extraction. To address class imbalance, SMOTE and ADASYN oversampling techniques are applied. Multiple classifiers including Random Forest, Gradient Boosting, XGBoost, and K-Nearest Neighbors are trained and evaluated. Experimental results show that Random Forest combined with SMOTE achieves the best balance between accuracy and fraud detection capability. The proposed approach provides a scalable and reliable solution for improving the safety and trustworthiness of online recruitment systems.*

Keywords*:* Fraud Detection, Machine Learning, Natural Language Processing, Online Recruitment, SMOTE, TF-IDF.

I. INTRODUCTION

Online recruitment portals have become a dominant medium for job search and talent acquisition due to their accessibility, efficiency, and global reach. However, the same openness has been exploited by malicious actors to publish fraudulent job postings that closely resemble legitimate advertisements. These fake postings often promise unrealistic salaries, remote work opportunities, or urgent hiring, thereby deceiving job seekers into sharing confidential information or making financial transactions.

Traditional rule-based or keyword-driven filtering systems are insufficient to detect such sophisticated fraud patterns. The linguistic similarity between real and fake job descriptions further complicates detection. Consequently, machine learning (ML) and natural language processing (NLP) techniques have emerged as effective solutions for identifying hidden patterns and anomalies within textual data.

This paper focuses on developing a machine learning framework that leverages NLP-based feature extraction and ensemble learning models to accurately classify job postings as genuine or fraudulent. Special emphasis is placed on addressing dataset imbalance, which is a critical challenge in fraud detection problems.

Recent studies have shown that machine learning and natural language processing techniques are effective in detecting online fraud and anomalous textual patterns [3], [4], [8].

II. LITERATURE SURVEY

Several studies have explored the application of machine learning for online recruitment fraud detection. Early research relied on traditional classifiers such as Naïve Bayes and Logistic Regression, which offered limited performance due to their inability to capture complex feature interactions. Recent studies demonstrate that ensemble models such as Random Forest and Gradient Boosting significantly improve detection accuracy.

DOI: 10.48175/IJARSCT-31064

Text preprocessing and feature extraction play a crucial role in fraud detection. TF-IDF has been widely adopted due to its ability to highlight discriminative terms within job descriptions. Additionally, class imbalance has been identified as a major issue, with techniques like SMOTE and ADASYN proving effective in enhancing minority class representation.

To address the issue of class imbalance in the dataset, synthetic oversampling techniques such as SMOTE and ADASYN were employed, as these methods have been shown to significantly improve classification performance in imbalanced learning scenarios [1], [2].

Despite these advancements, many existing approaches suffer from limited interpretability and lack scalability for real-world deployment. This study aims to bridge these gaps by integrating balanced learning with robust ensemble classifiers.

III. PROPOSED METHODOLOGY

The proposed system follows a structured workflow comprising data preprocessing, feature extraction, data balancing, model training, and evaluation.

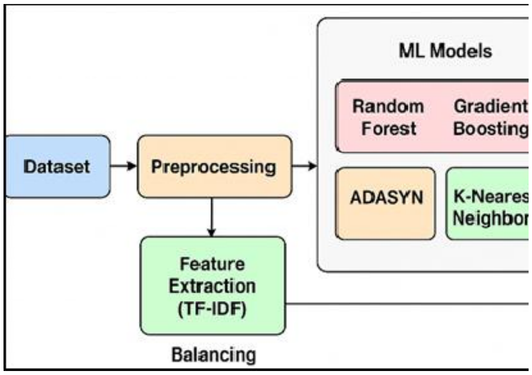

Figure 1: System Architecture

A. Dataset Description

The Employment Scam Aegean Dataset (EMSCAD) contains approximately 18,000 job postings labeled as fraudulent or legitimate. Each record includes attributes such as job title, company profile, job description, and requirements.

B. Data Preprocessing

Textual data is cleaned using tokenization, stopword removal, lowercasing, lemmatization, and removal of special characters. This step ensures noise-free and standardized input for feature extraction.

C. Feature Extraction

TF-IDF is used to convert textual data into numerical feature vectors. This representation captures the importance of words relative to the corpus and is effective in identifying deceptive linguistic patterns.

D. Handling Class Imbalance

To mitigate bias toward legitimate postings, SMOTE and ADASYN oversampling techniques are applied. These methods generate synthetic samples for the minority class, enabling balanced learning.

E. Machine Learning Models

The following classifiers are implemented:

- Random Forest
- Gradient Boosting

- XGBoost
- K-Nearest Neighbors

Table I: Comparative Performance of Machine Learning Models

Model	Balancing Technique	Accuracy (%)	Precision	Recall	F1-Score
Random Forest	SMOTE	98.2	0.94	0.76	0.84
Gradient Boosting	SMOTE	97.6	0.91	0.73	0.81
XGBoost	ADASYN	97.3	0.90	0.70	0.79
KNN	SMOTE	96.9	0.89	0.71	0.78

Random Forest with SMOTE achieves the highest accuracy and F1-score, indicating superior performance in detecting fraudulent job postings.

Ensemble-based classifiers were selected due to their robustness and superior generalization capability in handling nonlinear feature relationships [6], [7].

From the results, it is evident that Random Forest combined with SMOTE delivers the best overall performance, particularly in terms of F1-score and recall, which are critical for fraud detection tasks.

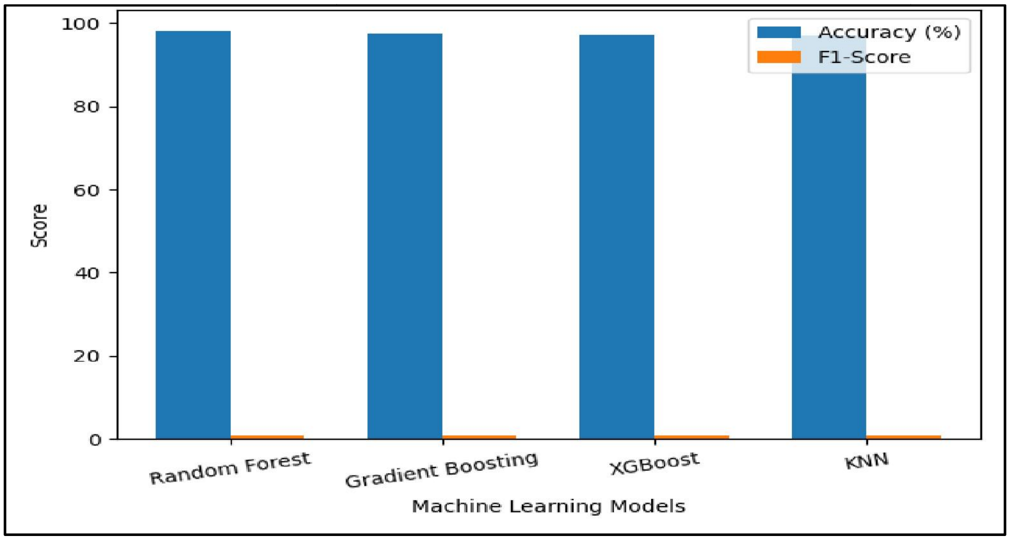

Figure 2: Comparison Graph

Figure 2 illustrates a comparative graph of model accuracy and F1-score. Ensemble-based models consistently outperform traditional classifiers due to their robustness and ability to capture nonlinear relationships. The results confirm that data balancing significantly improves fraud detection recall, thereby reducing false negatives.

IV. CONCLUSION

This paper presented a machine learning–based framework for detecting fraudulent online job postings using NLP and ensemble learning techniques. The integration of TF-IDF feature extraction with SMOTE-based data balancing and Random Forest classification yields superior detection performance. The proposed system offers a practical and scalable solution for enhancing security in online recruitment platforms.

Future work will focus on incorporating explainable AI techniques, real-time detection, and GUI-based deployment to improve usability and transparency.

DOI: 10.48175/IJARSCT-31064

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Prof. Ms. Swati S. Hinge for her valuable guidance, continuous support, and constructive suggestions throughout the course of this research work. Her insights and encouragement were instrumental in shaping this study.

The authors also thank the Department of Artificial Intelligence, Sanghavi College of Engineering for providing the necessary resources and academic environment to carry out this research

REFERENCES

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[2] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” in Proc. IEEE Int. Joint Conf. Neural Networks, 2008, pp. 1322–1328.

[3] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.

[4] C. Phua, V. Lee, K. Smith, and R. Gayler, “A Comprehensive Survey of Data Mining-based Fraud Detection,” Artificial Intelligence Review, Springer, 2010.

[5] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[6] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC, 2012.

[7] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” SN Computer Science, Springer, 2021.

[8] S. Saini and N. Kaur, “Online Job Fraud Detection Using Machine Learning,” International Journal of Computer Applications, vol. 176, no. 18, pp. 1–6, 2020.