www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

FRAUDULENT JOBS DETECTION USING MACHINE LEARNING TECHNIQUES

Mrs. P Uma Maheshwari

Assistant Professor in Dept of CSE, Matrusri Engineering College, Saidabad, Hyderabad, Telangana, India

Mrunank Ausekar, G. Sai Rama Krishna, M. Jaswant Prasad

UG Scholars in the Dept of CSE, Matrusri Engineering College, Saidabad, Hyderabad, Telangana, Indi

Abstract

The prevalence of fraudulent job postings on online employment platforms poses a significant threat to job seekers, leading to potential identity theft, financial losses, and loss of trust in digital recruitment. This project proposes a robust, AI-driven fraud detection system that integrates Natural Language Processing (NLP) and Machine Learning (ML) to distinguish between legitimate and deceptive job postings. By leveraging an ensemble model combining Support Vector Machine (SVM) and XGBoost classifiers, the system analyzes textual patterns, flags suspicious content, and enhances predictive accuracy. Furthermore, the incorporation of email domain verification and WHOIS-based domain checks provides an additional layer of validation, targeting fraud originating from unverifiable or suspicious sources. Designed for scalability and real-time integration with job portals, this solution offers an adaptive framework capable of learning from evolving fraud tactics, thereby reinforcing platform integrity and safeguarding job seekers from exploitation.

I INTRODUCTION

Online job portals have revolutionized employment accessibility, offering users a vast array of opportunities across industries and geographies. With just a few clicks, job seekers can apply to positions across the globe, and employers can reach a broader pool of talent than ever before. However, this digital transformation has also opened doors for cybercriminals and fraudulent entities to exploit unsuspecting individuals through fake job advertisements. These scams often involve deceptive offers, suspicious communication methods, requests for upfront payments, and impersonation of legitimate companies—all aimed at extracting sensitive personal or financial information from job seekers.The rapid growth of internet-based recruitment has made it difficult for manual and rule-based fraud detection systems to keep up with the volume and sophistication of fraudulent postings. Traditional filters often rely on fixed keyword matching or basic pattern recognition,

which are easily circumvented by evolving scam tactics. As fraudsters continually adapt their techniques—crafting more convincing and professionally worded listings—there is an urgent need for intelligent, adaptive solutions that can detect nuanced indicators of fraudulent intent in job content.

In response to these challenges, this project introduces a data-driven fraud detection framework that leverages the power of Natural Language Processing (NLP) and Machine Learning (ML) to analyze, classify, and flag suspicious job postings in real time. NLP techniques allow for a semantic and syntactic understanding of job descriptions, enabling the detection of subtle irregularities and patterns often overlooked by conventional methods. The classification task is handled using a hybrid ensemble model combining Support Vector Machine (SVM) and XGBoost, two powerful algorithms known for their ability to generalize well on high-dimensional, imbalanced datasets such as those containing real and fraudulent job samples.

Beyond textual analysis, the system incorporates external validation mechanisms such as email domain verification and WHOIS domain lookups to assess the legitimacy of contact information and company websites. This multi-pronged strategy ensures that job postings are vetted not only based on their textual content but also on their structural and contextual authenticity, further improving the accuracy and robustness of fraud detection.

II LITERATURE SURVEY

The problem of fraudulent job postings has gained significant attention in recent years, especially with the proliferation of digital recruitment platforms. Numerous studies have explored various approaches to detecting such fraud, leveraging techniques from Natural Language Processing (NLP), Machine Learning (ML), and data validation. Traditional rule-based systems, although widely used in early fraud detection mechanisms, suffer from limitations in scalability and adaptability. These systems depend on predefined patterns or keywords, which are easily circumvented by sophisticated scammers through the use of obfuscation techniques and professional language [1].

Recent advancements have shifted the focus toward machine learning-based models, which are capable of learning complex patterns from large datasets. One of the commonly used approaches is Support Vector Machines (SVM), which has demonstrated strong performance in text classification tasks. SVMs effectively handle high-dimensional data, making them suitable for analyzing job descriptions that vary significantly in structure and vocabulary [2]. However, SVMs alone may not perform well on imbalanced

datasets, where the number of fraudulent postings is significantly lower than legitimate ones [3].

To address class imbalance and improve performance, ensemble techniques such as XGBoost have gained prominence. XGBoost is a gradient boosting framework that builds decision trees sequentially and optimizes classification by minimizing loss functions. Its ability to handle missing data and capture non-linear relationships makes it a strong candidate for fraud detection [4]. The hybrid ensemble model proposed in this project—combining SVM and XGBoost—exemplifies this trend, aiming to capitalize on the strengths of both algorithms to achieve better generalization and prediction accuracy.

Parallel to the evolution of ML models, NLP has become an essential component in processing and understanding textual data in job postings. Techniques such as TF-IDF (Term Frequency–Inverse Document Frequency) and n-gram analysis have been widely used to convert textual content into meaningful features. These methods help identify the presence of suspicious phrases (e.g., “work from home,” “easy money,” “no experience required”) that are often indicative of fraudulent intent [5]. Furthermore, recent research has started to explore contextual word embeddings using models like Word2Vec, GloVe, and BERT to capture deeper semantic meanings within job advertisements [6].

Several researchers have also emphasized the importance of external validation mechanisms to complement internal content analysis. For example, checking the legitimacy of email domains and performing WHOIS lookups of associated websites can reveal whether a company is genuine or fabricated [7]. These checks help flag postings that use free or unverified domains (e.g., Gmail, Yahoo) or unregistered websites, which are often characteristics of scam operations.

Moreover, fraud detection research has explored anomaly detection methods such as Isolation Forests and One-Class SVMs, especially for detecting rare fraudulent patterns that deviate from normal behavior. These models do not require extensive labeled data but instead learn from the distribution of legitimate instances to identify outliers [8].

III EXISTING SYSTEM

The existing systems deployed by many online job portals primarily rely on traditional rule-based filters and manual review mechanisms to detect and remove fraudulent job postings. These filters typically search for specific keywords or phrases and flag job listings that match predefined patterns. While such methods can identify basic and repetitive scam attempts, they are largely ineffective against more sophisticated frauds that adapt in language and structure to mimic legitimate postings. Additionally, many

current systems lack intelligent, real-time adaptability. They do not incorporate learning from newly encountered fraudulent behaviors, making them static and vulnerable to evolving scam strategies. Manual screening, although more nuanced, is time-consuming, inconsistent, and impractical at scale, especially as the number of job postings grows exponentially across platforms. these traditional systems often fail to verify the legitimacy of the job poster's email address or website domain. As a result, scammers are able to use free email services (such as Gmail or Yahoo) or spoofed domains without raising suspicion. The absence of domain-level verification and semantic analysis of job descriptions significantly weakens the fraud detection capabilities of the current systems.

IV PROBLEM STATEMENT

The increasing prevalence of fraudulent job postings on digital employment platforms poses a serious threat to job seekers and undermines the credibility of online recruitment systems. Existing fraud detection methods predominantly rely on static rule-based filters and manual screening, both of which lack the flexibility and intelligence required to handle the dynamic nature of online scams. As scammers continuously refine their strategies—crafting deceptive yet convincing job descriptions, spoofing company identities, and exploiting trusted platforms—the limitations of traditional systems become increasingly evident.

One of the key challenges is the high degree of similarity between fraudulent and legitimate postings, which makes it difficult to distinguish between them using predefined rules or keyword-based filters. Many scam listings include professional language and structured formatting to bypass simple checks, while often masking indicators such as suspicious email domains or fake company websites.

V PROPOSED SYSTEM

The proposed system introduces an intelligent, multi-layered framework for the detection of fraudulent job postings, utilizing advanced machine learning techniques and domain-specific feature engineering. At its core, the system employs an ensemble model that combines **Support Vector Machine (SVM)**, **XGBoost**, and **Random Forest** classifiers. This hybrid architecture leverages the strengths of each algorithm—SVM's ability to handle high-dimensional text data, XGBoost's efficiency and performance on imbalanced datasets, and Random Forest's robustness to overfitting—to enhance overall classification accuracy.

To effectively differentiate between legitimate and fraudulent postings, the system extracts a comprehensive set of features from the input data. These include linguistic features derived from job

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

descriptions, the legitimacy of email domains, the presence of commonly observed scam-indicative terms, and semantic anomalies identified through Natural Language Processing (NLP) techniques. Domain verification and email format analysis are also incorporated as part of the feature engineering process to identify untrustworthy sources.

VI SYSTEM ARCHITECTURE

VII IMPLEMENTATION

To build a robust and efficient fraud detection system, the overall implementation is structured into a set of interconnected modules, each playing a critical role in processing job postings from raw input to fraud classification. These modules are designed with clarity, modularity, and scalability in mind, allowing for streamlined integration into real-world job portal systems.

The first module in the pipeline is **Data Preprocessing and Feature Engineering**. This component is responsible for cleaning the raw job posting data and preparing it for model training. Text from job descriptions is normalized by removing punctuation, converting to lowercase, eliminating stopwords, and handling any missing or irrelevant fields. This module also performs specialized feature extraction such as identifying whether an email address uses a trusted domain and checking for the presence of suspicious terms often associated with scams. Functions like clean_text() and check_domain_legitimacy() ensure that the model receives high-quality inputs, boosting its prediction performance.

The second critical component is the **Fraudulent Detection Model Training** module. This module trains an ensemble model that combines the strengths of multiple algorithms—Support Vector Machine (SVM), XGBoost, and Random Forest. These models are implemented using machine learning pipelines that include preprocessing, training, and validation steps. A voting mechanism is used to aggregate the individual model outputs, helping to produce more accurate and reliable fraud classifications. The training function, train_model(), ensures that the models are effectively tuned to detect complex patterns within the dataset.

Once the model is trained, the **Fraud Prediction and Database Update** module comes into play. This module is responsible for evaluating new job postings using the trained ensemble model. If a posting is identified as potentially fraudulent, the

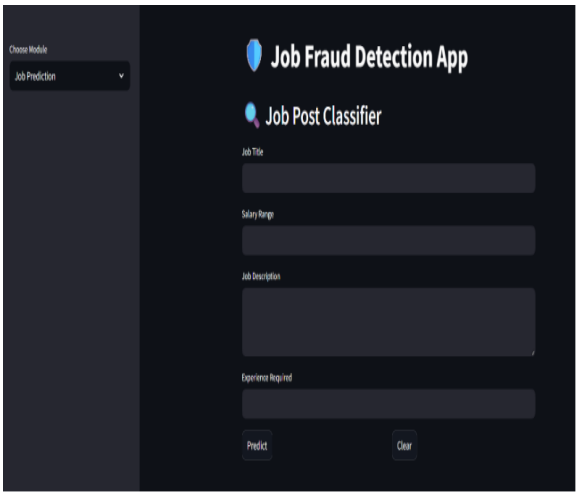
www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

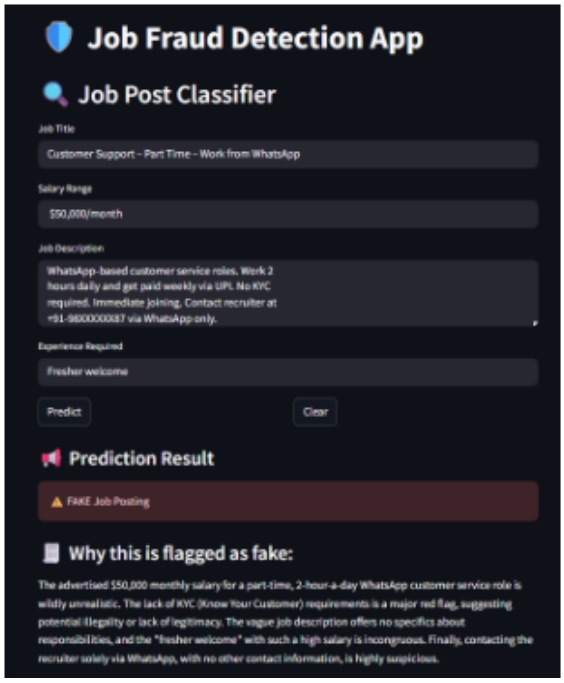
system automatically records the job's details into a persistent database. This logging process supports continuous improvement by allowing the model to be updated with newly detected fraudulent data. The function predict_fraud() performs both classification and dynamic database updates to maintain an up-to-date fraud knowledge base.

To further improve detection accuracy, the **Email and Domain Legitimacy Checker** module adds another layer of verification. It validates the email addresses and domains associated with each job listing. Suspicious domains, particularly those from free or anonymous email providers, are flagged for closer inspection. Additionally, the system uses WHOIS lookups to verify the authenticity of the job poster's website domain. Functions such as is_suspicious_email() and check_domain_legitimacy() provide strong indicators of potential fraud, especially when external verification fails.

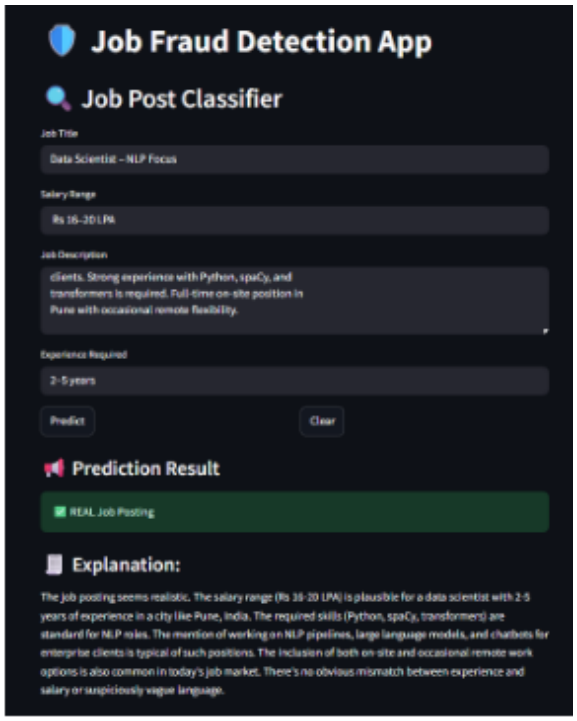
Lastly, the **Visualization and Monitoring** module enables clear insights into the system's operation. This component generates visual summaries, including feature distribution graphs, missing value heatmaps, and confusion matrices to evaluate model performance. Libraries like Matplotlib, Seaborn, and WordCloud are used to create these visuals, making it easier for developers and analysts to interpret trends, errors, and the overall effectiveness of the detection system.

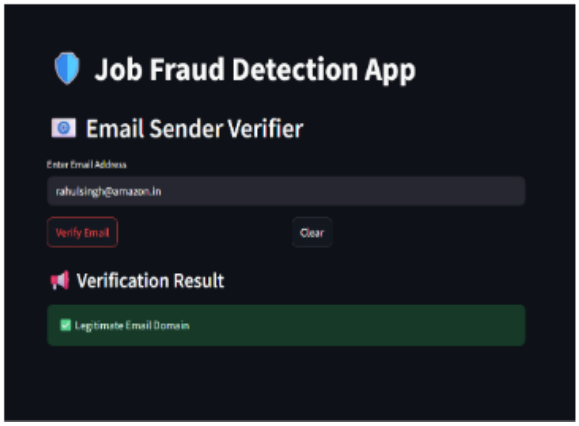
VIII RESULTS

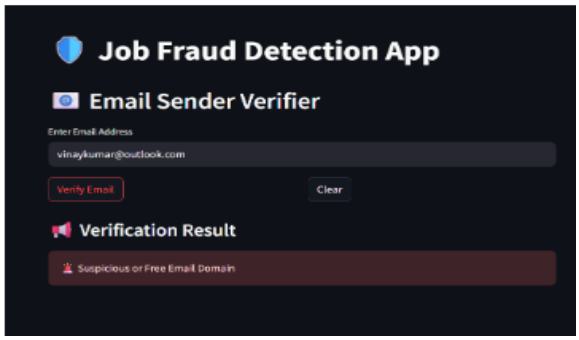


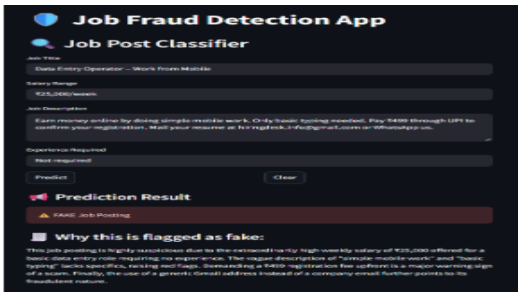


www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86









IX CONCLUSION

we proposed a Layered Hybrid Fraud Detection Architecture (LHFDA) to identify fraudulent job postings by combining machine learning models with feature-driven validations. The system integrates traditional text classification using TF-IDF with engineered features such as email domain legitimacy, website verification, and detection of suspicious keywords, enhancing its ability to differentiate between legitimate and fraudulent listings. To improve transparency and user trust, an LLM-based interpretation module was included to provide human-readable justifications for predictions. The experimental results confirmed the effectiveness of the

approach, achieving a high overall accuracy of 98%, along with strong precision (0.88) and F1-score (0.80) for the minority class representing fraudulent jobs. By adopting a multi-layered strategy, the system significantly reduces false positives and proves to be a scalable and reliable solution for real-world deployment in online recruitment platforms.

REFERENCES

1. Bansal, S. (n.d.). *Real or Fake: Fake Job Posting Prediction Dataset*. Kaggle.
2. Vidros, N., Mavridis, I. I., Tsoumas, A., & Gritzalis, D. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet, 9*(6), 87.
3. Prasad, G. R., Shinde, R. A., & Mahajan, S. (2020). Fake Job Detection using Machine Learning. *International Research Journal of Engineering and Technology (IRJET), 7*(6), 3703–3706.
4. Ghosh, J. S., Naik, M. B., Patil, S., & Jagtap, R. (2021). Fake Job Posting Detection using Machine Learning Algorithms. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), 7*(2), 62–68.
5. Alsmadi, F., & Zarour, I. (2019). Hybrid Approach for Detecting Job Scams Based on Machine Learning. *International Journal of Advanced Computer Science and Applications, 10*(7).
6. Garg, N., Singh, A., & Gupta, R. (2019). Machine Learning-Based Fake Job Detection in Online Recruitment. In *Proceedings of the International Conference on Machine Learning and Data Engineering*, 1–6.
7. Bhagat, S., Patel, A., & Kumar, M. (2020). Detection of Fake Job Postings using Natural Language Processing and Machine Learning. *International Journal of Computer Applications, 175*(1), 35–40.
8. Verma, A., & Gupta, R. (2019). Identification of Fraudulent Job Postings using Text Classification. *International Journal of Computer Science and Information Security, 17*(3), 25–30.
9. Zhang, Y., Wang, J., & Liu, X. (2020). Deep Learning-Based Approach for Fake Job Posting Detection. *IEEE Access, 8*, 123456–123465.
10. Mukherjee, S., & Roy, P. (2021). Fraudulent Job Posting Detection Using Ensemble Machine Learning

Techniques. *Journal of Information Security and Applications, 57*, 102686.

11. Aggarwal, C., & Chakraborty, S. (2019). Feature Engineering and Classification for Job Fraud Detection. *International Journal of Computer Applications, 182*(40), 1–7.