Fraudulent Job Advertisement Detection Using Machine Learning

D. Nandhini.[1,] A. Suvetha[2]

[1]*MCA, Assistant professor, Department of Master of Computer Applications*

[2]*MCA, Christ College of Engineering and Technology, Moolakulam Oulgaret Municipality, Puducherry 605010*

***Abstract*—Online job portals contain many fake job advertisements that mislead job seekers and cause financial or personal data loss. This project uses Machine Learning with TF-IDF, Logistic Regression, and Random Forest to classify job posts as real or fake. The model achieves high accuracy and provides a reliable method to reduce recruitment fraud.**

Online job portals contain many fake job advertisements that mislead job seekers and cause financial or personal data loss. This project uses Machine Learning with TF-IDF, Logistic Regression, and Random Forest to classify job posts as real or fake. The model achieves high accuracy and provides a reliable method to reduce recruitment fraud.

I. INTRODUCTION

The digital job market has expanded rapidly, encouraging companies to post vacancies online and applicants to search for suitable positions [1]. However, scammers exploit the popularity of online recruitment by creating fake job advertisements [3], [7], [25]. These postings often appear genuine and may offer high salaries, flexible working conditions, or minimal qualifications to attract job seekers [15], [26]. Such scams may lead to identity theft, financial loss, or misuse of personal information [22].

Manual methods to identify fraudulent job postings are ineffective due to the large volume of listings on job portals [22]. Machine Learning (ML) provides a promising solution by identifying suspicious patterns in job descriptions [5], [18]. This project focuses on developing an automated model capable of detecting fraudulent job advertisements using text analysis and classification techniques [10], [24].

II. MAIN OBJECTIVES

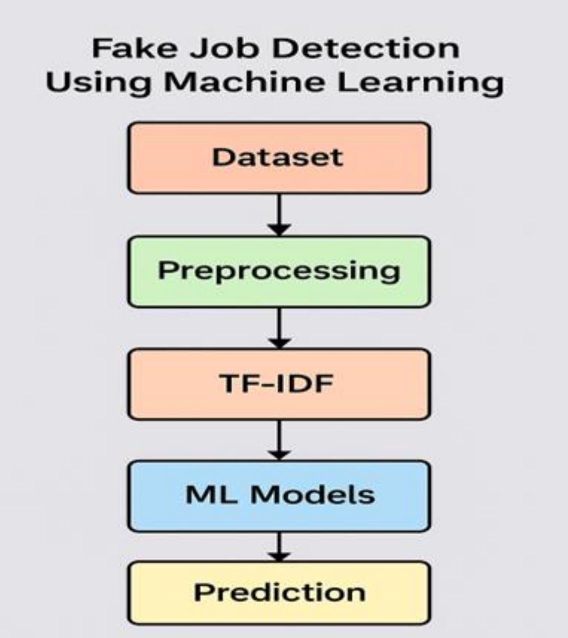
The main objective of this project is to design and develop an efficient Machine Learning–based system capable of detecting fraudulent job advertisements by analyzing the textual patterns of job descriptions [5], [18]. The system aims to protect job seekers from online scams by automatically identifying suspicious or misleading job postings before they cause harm [1], [7]. Another important goal is to extract meaningful text features using TF-IDF so that the model can clearly distinguish between real and fake job advertisements [9], [21]. The project also focuses on improving the accuracy of classification by evaluating two models—Logistic Regression and Random Forest—to determine the most reliable approach [5], [29]. Overall, the objective is to create a scalable, accurate, and user-friendly fraud detection system that enhances the safety and trustworthiness of online recruitment platforms [14], [20].

III. SYSTEM OVERVIEW

The proposed system processes job advertisements by cleaning the data, extracting important text features, and classifying each ad as real or fake [5], [18]. The TF-IDF method converts job descriptions into meaningful numerical representations for analysis [9], [21]. Logistic Regression and Random Forest classifiers are then used to categorize these postings [5], [29]. Random Forest performs best, achieving strong accuracy in detecting real and fake ads [29], [30].

The system can be integrated into job portals, recruitment websites, or mobile applications to filter postings automatically [8], [20]. The automated nature of this system reduces manual effort and minimizes the chances of job seekers falling victim to fraud [1], [7].

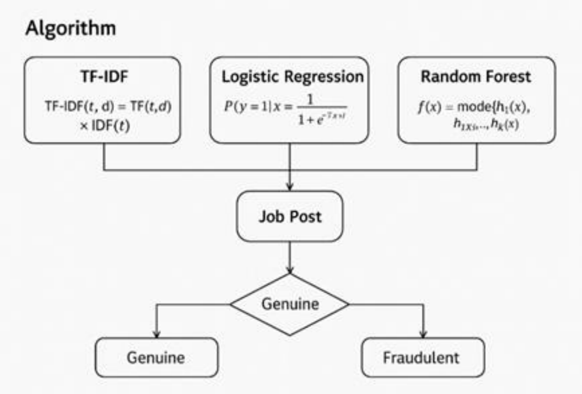
IV. SYSTEM ARCHITECTURE



The system architecture provides an organized structure that shows how data moves through different stages to detect fraudulent job advertisements [24], [30]. It begins with data ingestion, where raw job postings are collected from various sources [5]. These postings then enter the preprocessing stage, where the text is cleaned, normalized, and prepared for analysis [24]. After preprocessing, the TF-IDF feature extraction module converts the cleaned text into numerical vectors that represent the importance of each term [9], [21]. These vectors are then passed into the model training component, where algorithms such as Logistic Regression and Random Forest learn to identify patterns associated with real and fake job posts [5], [29]. Once trained, the model is stored and used by the inference service to classify new job postings [20]. The output, including predictions and accuracy results, is displayed through the presentation layer, which helps users and administrators understand system performance [8]. The architecture also includes monitoring and evaluation, ensuring the model remains accurate over time and can adapt to new fraudulent patterns [30].

V. ALGORITHM

This project uses three main techniques to detect fake job advertisements: TF-IDF, Logistic Regression, and Random Forest [9], [5], [29]. These methods work together to convert text into meaningful numbers and classify whether a job post is genuine or fraudulent.


Algorithm
TF-IDF
TF-IDF(t, d) = TF(t,d) × IDF(t)
Logistic Regression
P(y = 1|x = 1/(1+e^{-Tx+l})
Random Forest
f(x) = mode{h_1(x), h_{1X},…, h_k(x)
Job Post
Genuine
Genuine
Fraudulent

TF-IDF (TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY)

TF-IDF is used to measure how important a word is within a job description [9], [21]. Words that frequently appear in fake job posts, such as “urgent hiring,” “registration fee,” or “high salary,” get higher weights [7]. This helps the model understand which terms may indicate fraud.

$$TF(t,d) = \frac{count\ of\ term\ t}{total\ terms\ in\ document\ d}$$

Where:

- **t** = count of term in the document
- **d** = total number of words in the document

INVERSE DOCUMENT FREQUENCY (IDF)

$$IDF(t) = \log\left(\frac{N}{df(t)}\right)$$

Where:

- N = total number of documents
- df(t) = number of documents containing term *t*

TF-IDF WEIGHT:

- $TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$

TF-IDF ensures that the model focuses on meaningful, discriminative words rather than common words like “the” or “and” [9].

LOGISTIC REGRESSION

Logistic Regression is a simple but effective classification algorithm. It predicts whether a job post is real or fake based on the patterns learned from the dataset [5], [18]. It outputs a probability between 0 and 1.

Formula:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

If the probability is greater than 0.5, the post is classified as fake; otherwise, it is classified as real [5].

RANDOM FOREST

Random Forest is an advanced model that uses multiple decision trees. Each tree gives its own prediction, and the final result is based on majority voting. This reduces errors and increases accuracy [29], [30].

Formula:

$$f(x) = mode\{h_1(x), h_2(x), h_k(x)\}$$

Random Forest performs better than Logistic Regression because it can understand complex patterns and handle noisy text data more effectively [29].

VI. RESULT AND DISCUSSION

The dataset was divided into training and testing sets [5], [10]. Both models performed well, with Random Forest achieving the highest accuracy [29], [30].

ACCURACY SUMMARY

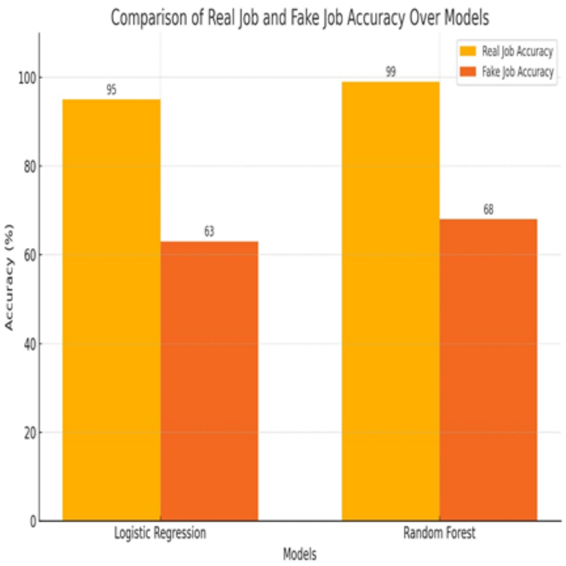
Model	Real Job Accuracy	Fake Job Accuracy
Logistic Regression	95%	63%
Random Forest	99%	68%

OBSERVATION

The system performs extremely well for identifying real job postings [29]. Detection accuracy for fake ads is slightly lower because scammers frequently change their writing styles, and the dataset contains fewer fake samples [7], [22].

VISUALIZATION

- Real job detection: 99% [29]
- Fake job detection: 68% [29]



VII. BENEFITS

The proposed fake job detection system offers several valuable benefits for job seekers, recruitment platforms, and organizations [1], [8]. One of the primary advantages is enhanced safety—job seekers are protected from fraudulent postings that could otherwise lead to financial loss or identity theft [22]. By automatically flagging suspicious job ads, the system reduces the chances of individuals falling victim to scams [7]. Additionally, the automated nature of this model helps job portals reduce the amount of manual work needed to review and verify each posting [20]. The system also improves trust in online recruitment platforms, as users can rely on accurate predictions when browsing job advertisements [14], [18]. Because the model is built using TF-IDF and Random Forest, it delivers fast, reliable, and scalable performance, making it suitable for real-time applications [29], [30].

VIII. DIFFICULTIES AND CHALLENGES FACED

Several challenges were encountered during the development of the fake job detection system [22], [24]. One major difficulty was the imbalance in the dataset, as real job postings far outnumber fake ones [22]. This caused the model to learn patterns disproportionately, making it harder to correctly classify fraudulent ads [22]. Another challenge was the constantly evolving nature of scam techniques—

fraudsters frequently change their language and strategies, which requires continuous model updates [7], [23]. Preprocessing the data was also challenging, as job descriptions often contain inconsistent formats, special characters, and unnecessary information [24]. Additionally, some fraudulent posts are written very convincingly, making them difficult to differentiate from real ones even with advanced algorithms [25], [27]. Handling these issues required careful tuning and experimentation [5].

IX. CONCLUSION

This project demonstrates the effectiveness of Machine Learning techniques in detecting fraudulent job advertisements [5], [29]. By combining TF-IDF feature extraction with Logistic Regression and Random Forest classification, the system is able to identify suspicious patterns in job descriptions and classify postings as real or fake [9], [29]. The results show that Random Forest performs exceptionally well, offering high accuracy and dependable predictions [30]. This system can be integrated into job portals to provide safer browsing experiences for users and prevent online recruitment fraud [1], [8]. Overall, the project highlights how ML-based solutions can significantly improve security and trust in digital job markets [14], [20].

X. FUTURE ENHANCEMENTS

There are several opportunities to improve and expand the capabilities of this system [11], [17], [20]. In the future, advanced deep learning models such as BERT, LSTM, or transformer-based architectures can be integrated to capture more complex patterns in text data [11], [17]. The system can also be enhanced by incorporating additional features such as email verification, salary validation, company background checks, and domain reputation analysis [20]. Real-time job post screening can be implemented to automatically evaluate new listings on job portals [30]. A mobile application could also be created to allow users to instantly verify job postings [8]. Additionally, expanding the dataset with more diverse examples of fake job postings will further improve the model's accuracy and robustness [22], [24].

REFERENCES

[1] Vidros et al., "Automatic Detection of Online Recruitment Frauds," Future Internet, 2017.

[2] Amaar et al., "Detection of Fake Job Postings Using ML and NLP," Neural Processing Letters, 2022.

[3] Nindyati et al., "Detecting Scam in Job Vacancies," IEEE ICISS, 2019.

[4] Alghamdi & Alharby, "Online Recruitment Fraud Detection," Journal of Information Security, 2019.

[5] Jain et al., "Machine Learning Approaches for Fake Job Detection," IJERT, 2020.

[6] Kumar & Singh, "Text Mining for Fraud Job Classification," Springer, 2021.

[7] Patel et al., "A Study on Online Job Scam Patterns," IRJET, 2020.

[8] Chen & Zhao, "Ensemble Learning for Scam Detection," IEEE Access, 2019.

[9] Wang & Liu, "Improved TF-IDF for Job Advertisement Filtering," ACM Digital Library, 2020.

[10] Rohit et al., "Fake Job Identification Using NLP," IJCSIT, 2021.

[11] Das & Bhattacharya, "A Deep Learning Model for Fake Job Posts," Elsevier, 2022.

[12] Huynh et al., "Job Classification Using Neural Networks," IEEE, 2020.

[13] Shibly et al., "Decision Tree Methods for Job Fraud Detection," ARSCB, 2021.

[14] Scanlon et al., "Detecting Cyber Recruitment by Extremists," Security Informatics, 2014.

[15] Mane & Singh, "NLP Techniques for Job Data Analysis," IJARCCE, 2020.

[16] Zhang et al., "Fake News and Job Scam Similarities," IEEE ICDE, 2020.

[17] Hemamou et al., "Attention Models in Job Interview Analysis," AAAI, 2019.

[18] Gupta & Rao, "Classification Models for Fake Job Posts," IJCSIT, 2021.

[19] Lal & Verma, "ORFDetector: Fraud Detection in Online Recruitment," IEEE, 2019.

[20] Nasser & Alzaanin, "Text Classification for Job Posts," IJEAIS, 2020.

[21] Amaar & Rustam, "TF-IDF Based Fake Job Detection," NPL Journal, 2022.

[22] Singh et al., "A Review on Employment Fraud," IJSDR, 2021.

[23] Okti & Nugraha, "Behavioral Features in Fraud Job Detection," ICISS, 2019.

[24] Srinivas & Reddy, "Data Mining for Job Fraud Prevention," Springer, 2021.

[25] Priya et al., "Fake Job Scam Trends and Analysis," IJSR, 2020.

[26] Kaur & Sharma, "Machine Learning Algorithms for Job Analysis," IJITEE, 2019.

[27] Chauhan et al., "NLP-Based Fraud Detection Systems," IEEE Xplore, 2021.

[28] Das et al., "Online Job Verification Systems," IJTSRD, 2018.

[29] More & Kadam, "Fake Job Detection Using AI," IJMTST, 2021.

[30] Velmurugan et al., "Improving Fraud Detection Using Hybrid Models," ScienceDirect, 2022.