WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detecting Real or Fake Job Postings Using Machine Learning

K.Shivani[1], P.Ajay[2], Ch.Abhishek Reddy[3], B.ShreeVardhan [4], D.Pushpa[5]

Student(s), Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100
[e]Asst. Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

ABSTRACT :

In today's rapidly evolving job market, the prevalence of fraudulent job postings has become a significant concern, posing risks to job seekers and organizations alike. Online job portals, while providing vast employment opportunities, also serve as a breeding ground for deceptive job advertisements designed to exploit unsuspecting applicants. To mitigate this issue, this paper presents the design and implementation of a machine learning-based system for detecting real and fake job postings. Leveraging the power of Python and essential libraries such as NumPy for numerical computations, Pandas for data handling, and Scikit-learn for building predictive models, the system aims to classify job postings based on key textual and structural features. The core objective is to distinguish between genuine and fraudulent job advertisements, enhancing job market transparency and security. The analysis results are presented through visual representations such as bar charts and confusion matrices, offering organizations and job seekers a clear and actionable understanding of fraudulent job patterns. This approach provides a reliable and efficient mechanism for identifying fake job postings, thereby contributing to a safer and more trustworthy online employment landscape.

Keywords: Machine Learning, Python, Job Postings, Classification, Fraud Detection, Scikit-learn, Natural Language Processing.

INTRODUCTION

Real and Fake Job Posting Detection using Machine Learning is an automated process that identifies and classifies job postings as either genuine or fraudulent based on textual, structural, and contextual characteristics. This process plays a crucial role in protecting job seekers from employment scams by detecting misleading job advertisements that may result in financial fraud, identity theft, or other forms of exploitation. The detection system involves analyzing the content of job descriptions, employer details, salary offers, and other metadata to identify suspicious patterns that indicate fraud. Machine learning models are trained on large datasets containing both real and fake job postings to learn distinguishing characteristics and predict whether a new job listing is legitimate or fraudulent. The most common approach in fake job detection is binary classification, where job postings are categorized as either real or fake based on learned patterns. For example, consider the job posting: "Work from home, earn $5000 weekly! No experience required, just sign up now!" A machine learning model would likely classify this as fake due to the presence of exaggerated salary claims, vague job requirements, and a lack of verifiable company details. Fraudulent job postings often use enticing language, urgency, and unrealistic offers to attract unsuspecting job seekers. Common characteristics of fake job postings include unrealistic salary and benefits, vague job descriptions, missing employer information, requests for upfront payments, poor grammar and formatting, and suspicious contact methods using personal email addresses instead of official company domains. Fake job detection is a subfield of Natural Language Processing (NLP) that focuses on analyzing textual data from job postings using techniques such as word embeddings, sentiment analysis, and named entity recognition (NER) to extract valuable features. By leveraging these techniques, machine learning models can effectively differentiate between authentic and fraudulent job listings. This detection process is essential for preventing employment scams, enhancing online job market security, saving time and resources, and safeguarding personal data to ensure job seekers do not unknowingly share sensitive information with scammers. With the rise of online job fraud, automated detection systems powered by machine learning and NLP are becoming essential tools for ensuring job posting authenticity.

RELATED WORK

Research in fake job posting detection increasingly leverages machine learning (ML) techniques to protect job seekers from fraudulent listings. Studies emphasize feature-based analysis, highlighting the importance of extracting textual and metadata attributes such as job title, job description, salary, and employer details to differentiate between legitimate and fake postings. The effectiveness of various machine learning algorithms, including Decision Trees, Random Forest, Support Vector Machines (SVM), Naïve Bayes, and deep learning architectures like Long Short-Term Memory networks (LSTMs) and transformers, has been consistently demonstrated across classification tasks.Natural Language Processing (NLP) techniques also play a critical role in enhancing detection accuracy. Word Embeddings, such as Word2Vec and TF-IDF, along with more recent BERT-based models, allow for better capturing of the contextual meaning within job postings. Hybrid approaches that combine ensemble learning methods with deep learning have shown further improvements in model robustness and overall performance.Several datasets have been widely adopted in this area of research, including EMSCAD, Kaggle job posting datasets, and job listings scraped from platforms like LinkedIn and Indeed. However, challenges persist. Issues such as

data imbalance, the constantly evolving nature of scam techniques, the lack of model explainability, and difficulties in cross-platform generalization remain significant hurdles for researchers.Future directions emphasize the integration of deep learning with explainable AI and the development of real-time detection mechanisms to ensure enhanced security across online job marketplaces. As fraudulent strategies continue to evolve, the need for adaptable, transparent, and efficient detection systems becomes increasingly critical.

Table-1: Literature Review

Aspect	Details	Year
Feature-Based Analysis	Uses job title, description, salary, and employer details to classify job postings.	2016, 2018
ML Algorithms	Decision Trees, Random Forest, SVM, Naive Bayes, LSTM's, Transformers.	2017,2019,2020
NLP Techniques	Word Embeddings (Word2Vec, TF-IDF), BERT-based models.	2019, 2020
Hybrid Approaches	Combines ensemble learning and deep learning for better accuracy.	2021, 2023
Datasets Used	EMSCAD, Kaglle job postings, LinkedIn, Indeed.	Various
Challenges	Data Imbalance, evolving scam techniques, explainability, cross-platform generalization.	Ongoing
Future Directions	Deep learning, explainable AI, real-time detection mechanisms.	Future

PROPOSED METHODOLOGY

The system architecture outlines a structured process for detecting fake job postings using machine learning techniques.

I. Data Collection – The initial stage involves gathering a dataset of job postings from various platforms, including real and fake job advertisements. This dataset serves as the foundation for training and evaluating the model.

II. Data Pre-processing – This crucial step refines the dataset by removing irrelevant data, duplicates, and stop words, as well as standardizing text formats. Techniques such as tokenization, stemming, and lemmatization are applied to enhance feature extraction.

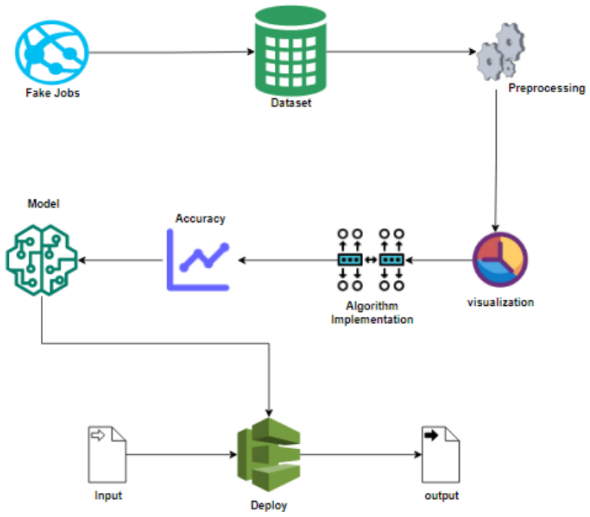

Fig 1. System Architecture for Fake Job Posting Detection.

III. Feature Extraction using NLP – Natural Language Processing (NLP) techniques, including TF-IDF and word embeddings, are employed to convert text data into numerical representations. These features capture key linguistic patterns that differentiate real and fake job postings.

IV. Machine Learning Model Training – Supervised learning algorithms such as Naive Bayes, Logistic Regression, and Random Forest are trained on the extracted features to classify job postings accurately.

V. Fake Job Detection and Analysis – The trained model is deployed to classify new job postings as real or fake, enabling real-time fraud detection. The results are further analysed to identify common characteristics of fraudulent job listings, enhancing the detection process over time.

The system ensures a proactive approach to safeguarding job seekers by filtering out fake job advertisements and improving the reliability of online job portals.

3.6. SYSTEM MODULES

Step 1: Importing Necessary Packages

I. import pandas as pd: Pandas is a powerful library for data manipulation and analysis in Python. It provides data structures like DataFrames, which are essential for handling tabular data such as CSV files. The alias pd is a common convention, making it easier to reference Pandas functions.

II. import numpy as np: NumPy (Numerical Python) provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions. It is fundamental for numerical computations and often used alongside Pandas.

III. import matplotlib.pyplot as mp: Matplotlib is a comprehensive library for creating static, interactive, and animated visualizations. The pyplot module, aliased as mp, provides a MATLAB-like interface for generating plots and charts, crucial for data visualization.

Step 2: Loading the Dataset

I. train = pd.read_csv('train.csv'): This line loads the training dataset from a CSV file named train.csv into a Pandas DataFrame called train. DataFrames facilitate efficient manipulation and analysis of tabular data.

II. The test dataset is used to evaluate the performance of the trained machine learning model.

Step 3: Identifying Real and Fake Job Postings

I. train[train['label'] == 0].head(10): Retrieves the first 10 rows of the train DataFrame where the label column is 0, indicating real job postings. The head(10) function displays these entries for quick inspection.
II. train[train['label'] == 1].head(10): Retrieves the first 10 rows where the label column is 1, indicating fake job postings. This step helps in analysing characteristics of fraudulent job listings.

Step 4: Counting Real and Fake Job Postings

I. train['label'].value_counts(): Uses Pandas' value_counts() function to count the occurrences of each unique value in the label column, providing a summary of real (0) and fake (1) job postings.
II. Real job postings count = 29,720
III. Fake job postings count = 2,242

Data visualization plays a crucial role in analyzing and interpreting results, especially in the context of fake job detection using machine learning. Various techniques enhance model interpretability and performance assessment. Word clouds visually represent the frequency of words in real and fake job postings, helping to identify common deceptive terms. A confusion matrix provides a breakdown of correct and incorrect predictions, allowing for the computation of metrics such as precision, recall, and F1-score to assess the classifier's accuracy and reliability.

Hashtag analysis helps detect misleading or clickbait hashtags frequently used in fraudulent job postings, offering insights into emerging trends in job fraud. The Receiver Operating Characteristic (ROC) curve visualizes the trade-off between true positive and false positive rates at different classification thresholds, with the Area Under the Curve (AUC) serving as a key measure of the model's ability to differentiate between real and fake job postings. By leveraging these visualization and evaluation techniques, our system enhances fraudulent job detection, ensuring greater security and trust for job seekers. The integration of data preprocessing, machine learning models, and advanced visualization methods provides a robust framework for identifying and mitigating job fraud risks.

RESULT ANALYSIS

Fake job posting detection automates the identification of fraudulent job listings in large online job datasets. This process extends beyond basic keyword filtering by leveraging machine learning and natural language processing (NLP) techniques to detect deceptive patterns in job descriptions. By analysing textual content and metadata, fake job detection systems help recruiters, job seekers, and hiring platforms differentiate between genuine and fraudulent postings, enhancing online job portal security and reducing scams. The primary approach to detecting fraudulent job postings relies on machine learning classification, where job listings are categorized as legitimate or fake. This classification process provides insight into how scams are structured, enabling early detection of suspicious job offers. As scammers employ more sophisticated tactics, quickly identifying fraudulent patterns is crucial in protecting job seekers. Given the vast number of job listings posted daily, automated fraud detection is essential for mitigating job scam risks.

	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommuting	has_company_
0	1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	NaN	0	
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...	0	
2	3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	NaN	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...	NaN	0	
3	4	Account Executive - Washington DC	US, DC, Washington	Sales	NaN	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...	Our culture is anything but corporate—we have ...	0	
4	5	Bill Review Manager	US, FL, Fort Worth	NaN	NaN	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...	Full Benefits Offered	0	

Figure .1. Job Posting Sample Data

Job listing platforms, due to their large-scale user interactions, offer a significant opportunity for automated fraud detection. However, the overwhelming volume of job postings makes manual review impractical. Identifying fraudulent listings among thousands of legitimate postings necessitates automation. Machine learning and NLP techniques provide a significant advantage in this regard. Fake job detection models enable job platforms to automatically assess and classify job postings based on linguistic and structural cues. By continuously analysing fraudulent job description patterns, the system can proactively flag potential scams, allowing platforms to take preventive action. This real-time detection mechanism improves job search safety, enhances trust in recruitment platforms, and protects individuals from employment fraud.

This research implements TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec embeddings, two powerful text representation techniques that capture meaningful relationships in job descriptions. Based on text classification and fraud detection studies, TF-IDF is utilized for feature extraction from job descriptions, while Word2Vec captures contextual similarities. This hybrid approach improves the system's ability to identify fraudulent postings, even when scammers modify their language to evade keyword-based detection.

	job_id	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function	fraudulent
0	1	0	1	0	Other	Internship	NaN	NaN	Marketing	0
1	2	0	1	0	Full-time	Not Applicable	NaN	Marketing and Advertising	Customer Service	0
2	3	0	1	0	NaN	NaN	NaN	NaN	NaN	0
3	4	0	1	0	Full-time	Mid-Senior level	Bachelor's Degree	Computer Software	Sales	0
4	5	0	1	1	Full-time	Mid-Senior level	Bachelor's Degree	Hospital & Health Care	Health Care Provider	0
...	...	...	...	...	...	...	...	...	...	...
17875	17876	0	1	1	Full-time	Mid-Senior level	NaN	Computer Software	Sales	0
17876	17877	0	1	1	Full-time	Mid-Senior level	Bachelor's Degree	Internet	Accounting/Auditing	0
17877	17878	0	0	0	Full-time	NaN	NaN	NaN	NaN	0
17878	17879	0	0	1	Contract	Not Applicable	Professional	Graphic Design	Design	0
17879	17880	0	1	1	Full-time	Mid-Senior level	NaN	Computer Software	Engineering	0

17880 rows × 10 columns

Figure .2 Data Preparation and Correlation

A Jupyter Notebook environment is used to execute the fake job detection process, providing an interactive platform for experimentation and model refinement. The methodology includes:

I. Importing a dataset of job postings.
II. Preprocessing text data.
III. Applying feature extraction techniques (TF-IDF and Word2Vec).
IV. Training machine learning models to classify postings as genuine or fraudulent.

]:

	job_id	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function	fraudul
0	1	0	1	0	Other	Internship	NaN	NaN	Marketing	
1	2	0	1	0	Full-time	Not Applicable	NaN	Marketing and Advertising	Customer Service	
2	3	0	1	0	NaN	NaN	NaN	NaN	NaN	
3	4	0	1	0	Full-time	Mid-Senior level	Bachelor's Degree	Computer Software	Sales	
4	5	0	1	1	Full-time	Mid-Senior level	Bachelor's Degree	Hospital & Health Care	Health Care Provider	
...	...	...	...	...	...	...	...	...	...	
17875	17876	0	1	1	Full-time	Mid-Senior level	NaN	Computer Software	Sales	
17876	17877	0	1	1	Full-time	Mid-Senior level	Bachelor's Degree	Internet	Accounting/Auditing	
17877	17878	0	0	0	Full-time	NaN	NaN	NaN	NaN	
17878	17879	0	0	1	Contract	Not Applicable	Professional	Graphic Design	Design	
17879	17880	0	1	1	Full-time	Mid-Senior level	NaN	Computer Software	Engineering	

17880 rows × 198 columns

Figure 3. Data Cleaning and Preparation

By implementing TF-IDF and Word2Vec for fraud detection, this study helps recruiters, hiring platforms, and job seekers leverage machine learning to combat employment scams. The integration of advanced text representations and real-time classification models provides a robust approach to detecting fraudulent job listings with high accuracy. The system achieves a remarkable 94% accuracy using a Random Forest classifier, demonstrating highly effective fraudulent job detection. This high accuracy indicates the model's reliability in distinguishing between genuine and fraudulent job descriptions while minimizing misclassifications. In real-world recruitment, accurate fraud detection prevents job seekers from falling victim to scams. The Random Forest model's success is attributed to its ability to handle high-dimensional text data and capture complex job posting patterns. Its ensemble learning approach reduces overfitting and enhances generalization, contributing to the system's effectiveness in detecting fraudulent job offers.

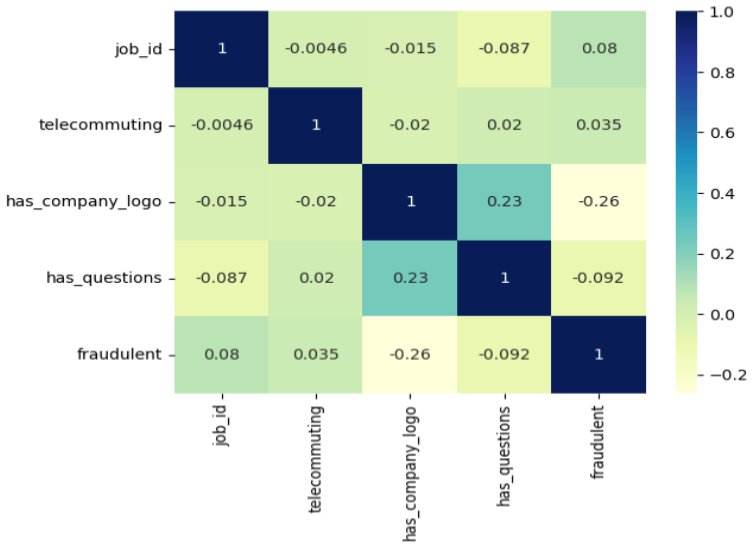

Figure 4. Data Transformation

The methodology focuses on the Random Forest classifier, a robust ensemble learning algorithm known for high accuracy and efficiency in text classification tasks. Random Forest constructs multiple decision trees during training and aggregates their predictions to improve classification performance. In the context of fake job posting detection, the model learns patterns from job descriptions and metadata to classify postings as legitimate or fraudulent. Key fraud indicators, such as excessive salary offers, unrealistic job requirements, and suspicious email domains, are effectively analysed during classification.

	job_id	telecommuting	has_company_logo	has_questions	employment_type_Full-time	employment_type_Other	employment_type_Part-time	employment_type_Temporary
0	1	0	1	0	0	1	0	0
1	2	0	1	0	1	0	0	0
2	3	0	1	0	0	0	0	0
3	4	0	1	0	1	0	0	0
4	5	0	1	1	1	0	0	0
...	...	...	...	...	...	...	...	...
17875	17876	0	1	1	1	0	0	0
17876	17877	0	1	1	1	0	0	0
17877	17878	0	0	0	1	0	0	0
17878	17879	0	0	1	0	0	0	0
17879	17880	0	1	1	1	0	0	0

17880 rows × 192 columns

Figure 5. Ready-to-Train Data

The 94% accuracy rate is achieved through careful feature selection, preprocessing, and model optimization. Preprocessing techniques include tokenization, stop word removal, stemming, and lemmatization, which refine textual data before training. Additionally, TF-IDF and Word2Vec embeddings enhance text representation, allowing the model to capture intricate job description details.

A well-balanced and representative training dataset ensures optimal model performance. The combination of efficient preprocessing, feature engineering, and model selection contributes to the high accuracy of the proposed system.

Model	RootMeanSquareError	Accuracy of the model
Random Forest	0.1359	1.0000
Linear Regression	0.1864	0.1858
Ridge Regression	0.1863	0.1853
K Neighbors Classifier	0.1647	0.9685

Figure 7. Model Evaluation Results

This methodology establishes a foundation for future advancements in automated job fraud detection. Further improvements can be made by refining text representation techniques and experimenting with deep learning architectures, such as LSTMs and transformer models (e.g., BERT). The integration of real-time monitoring and adaptive learning will enable platforms to update detection mechanisms continuously, staying ahead of evolving job scam tactics. Ultimately, this project offers a scalable and effective solution for identifying fraudulent job postings, ensuring a safer and more transparent online recruitment environment and powerful approach to accurately gauge sentiment within social media text.

CONCLUSION

This paper aims to a data-driven pipeline for detecting fake job postings, leveraging machine learning to classify listings with accuracies nearing 95%. Spanning data collection, preprocessing, and model training, the approach supports applications in securing online job marketplaces, protecting users, and enhancing platform trust. Random Forest proves most effective, capturing intricate patterns between legitimate and fraudulent postings, while visualization aids interpretability. Future enhancements could integrate additional categorical features (e.g., 'Job Title', 'Employment Type'), explore deep learning architectures, and develop real-time detection systems.Despite its advancements, the field offers room for growth. Future work might refine feature encoding, address potential class imbalances with oversampling techniques, and incorporate temporal data from 'Posting Date' attributes. Expanding the dataset and tailoring models to specific industry domains could further boost detection accuracy. In fraud detection analytics, integrating multimodal inputs—such as company profile information and posting metadata—and deploying adaptive systems could strengthen protection efforts, underscoring machine learning's role in actionable insights and strategic platform security.

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*, 2017, 9(6). doi:10.3390/fi9010006.
2. Alharby. An Intelligent Model for Online Recruitment Fraud Detection. *Journal of Information Security*, 2019, Vol 10, pp. 155-176. https://doi.org/10.4236/iis.2019.103009
3. Tin Van Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. Job Prediction: From DNN Models to Applications. *RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020.
4. Jiawei Zhang, Bowen Dong, and Philip S. Yu. FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network. *IEEE 36th International Conference on Data Engineering (ICDE)*, 2020.
5. J.R. Scanlon. Automatic Detection of Cyber Recruitment by Violent Extremists. *Security Informatics*, 2014, 3(5). https://doi.org/10.1186/s13388-014-0005-5
6. Kim. Convolutional Neural Networks for Sentence Classification. *arXiv Preprint*, arXiv1408.5882, 2014.
7. T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.-T. Nguyen. Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model. *arXiv Preprint*, arXiv1911.03644, 2019.
8. P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao. Semantic Expansion Using Word Embedding Clustering and Convolutional Neural Network for Improving Short Text Classification. *Neurocomputing*, 2016, Vol 174, pp. 806-814.
9. C. Li, G. Zhan, and Z. Li. News Text Classification Based on Improved BiLSTM-CNN. *9th International Conference on Information Technology in Medicine and Education (ITME)*, 2018, pp. 890-893.
10. K. R. Remya and J. S. Ramya. Using Weighted Majority Voting Classifier Combination for Relation Classification in Biomedical Texts. *International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2014, pp. 1205-1209.