WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Fake Job Detection and Analysis Using Machine Learning Algorithms

G. Vinod Kumar [1], *Mr. K. Srinivas Rao* [2]

[1]PG Scholar, Dept. of MCA, Aurora Deemed to Be University, Hyderabad, Telangana, India.[1]

[2]Assistant Professor, School of Informatics, Aurora Deemed to Be University, Hyderabad, Telangana, India.2

ABSTRACT

With the rapid increase of online job platforms, fraudulent job postings have also grown, targeting vulnerable job seekers. This paper proposes a machine learning-based system to classify job advertisements as real or fake using the EMSCAD dataset. Various supervised learning models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naive Bayes were evaluated based on their ability to identify scam patterns from textual job descriptions. The Random Forest classifier achieved the highest performance with an accuracy of 96.8%. In addition to model training and evaluation, a web application was developed using Flask to allow users to upload or paste job descriptions for real-time scam detection. Our approach emphasizes interpretability, feature selection, and real-time classification support for job portals and recruitment platforms.

Keywords: Fake Job Detection, EMSCAD, Machine Learning, Random Forest, Text Classification, Scam Prediction, Employment fraud, Flask, Web application

1. Introduction

Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF). In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by job-seekers. However, this intention may be one type of scam by fraudsters who offer employment to job-seekers in terms of taking money from them. Fraudulent job advertisements can be posted against a reputed company for violating their credibility. These fraudulent job post detections draw good attention for obtaining an automated tool for identifying fake jobs and reporting them to people for avoiding application for such jobs.

For this purpose, a rule-based approach is applied which employs classification based on predefined red flags for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job postings, a rule-based supervised learning algorithm is considered initially. A classifier maps input variables to target classes by considering training data. The classifier addressed in the paper for identifying fake job posts from the others is described briefly. This classifier-based prediction may be broadly categorized into Rule-Based Prediction and Potential ML Enhancements.

A. Rule-Based Prediction

The rule-based classifier is a supervised classification tool that exploits predefined rules based on common scam indicators. The decision made by this classifier is effective in practice even if the rules are simple. This classifier obtains promising results in scenarios where features are independent or functionally dependent. The accuracy of this classifier is not solely related to feature dependencies but to the information loss due to rule assumptions.

B. Potential ML Enhancements

Multi-layer perceptron can be used as a supervised classification tool by incorporating optimized training parameters. For a given problem, the number of hidden layers in a multilayer perceptron and the number of nodes in each layer can differ. The decision of choosing the parameters depends on the training data and the network architecture. K-Nearest Neighbor Classifiers, often known as lazy learners, identifies objects based on closest proximity of training examples in the feature space. The classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on choosing the appropriate value of k. A Decision Tree (DT) is a classifier that exemplifies the use of tree-like structure. It gains knowledge on classification. Each target class is denoted as a leaf node of DT and non-leaf nodes of DT are used as a decision node that indicates certain test. The outcomes of those tests are identified by either of the branches of that decision node. Starting from the beginning at the root this tree is going through it until a leaf node is reached. It is the way of obtaining classification result from a decision tree.

Decision tree learning is an approach that has been applied to spam filtering. This can be useful for forecasting the goal based on some criterion by implementing and training this model.

2. Literature Review

Previous studies highlight the use of natural language processing (NLP) techniques and machine learning classifiers to distinguish fake job listings. Some notable works include:

- Ray and Khan (2022) applied NLP and ML techniques to detect fake jobs using TF-IDF and classification models.
- Rao et al. (2021) developed a text classification framework using traditional algorithms like SVM and Logistic Regression.
- EMSCAD dataset was introduced as a benchmark for evaluating such systems, offering over 18,000 real and fake job descriptions.

However, many approaches lack explainability or real-time deployment capability. Our study aims to fill this gap.

3. Methodology

- Dataset: EMSCAD dataset from Kaggle, labeled with ‘fraudulent’ and ‘real’ classes.
- **Data Preprocessing:**
    - Removal of stopwords, punctuation, and special characters
    - Text normalization using lowercase conversion
    - Tokenization and vectorization using TF-IDF
- **Feature Selection:**
    - Important textual features: title, location, description, company profile, requirements
    - Correlation matrix used to reduce feature redundancy
- **Models Evaluated:**
    - Logistic Regression
    - Random Forest
    - Naive Bayes
    - Support Vector Machine (SVM)

The dataset from Kaggle [13] contains 17,880 job posts with attributes like title, description, and fraudulent. Preprocessing includes missing value removal, stop-word elimination, irrelevant attribute removal, and extra space removal.

The rule-based classifier uses a dictionary to detect red flags:

```
red_flags = {
    "urgent hiring": "The job post mentions 'urgent hiring', which is often used in scams to create a sense of urgency.",
    "no experience needed": "The job post claims 'no experience needed', but may offer an unrealistically high salary or require payment to apply.",
    "pay to apply": "The job post asks for payment to apply, which is a clear sign of a scam.",
    "immediate hiring": "The job post mentions 'immediate hiring', which can be a tactic to rush applicants without proper vetting."
}
```

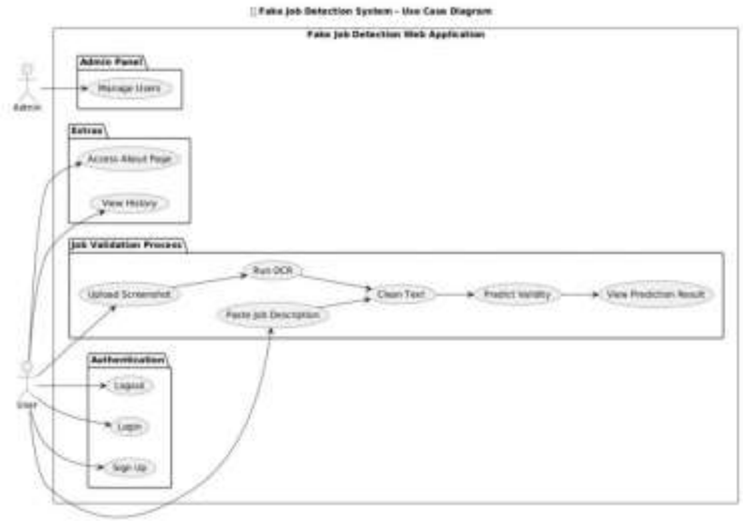
OCR preprocessing enhances image input:

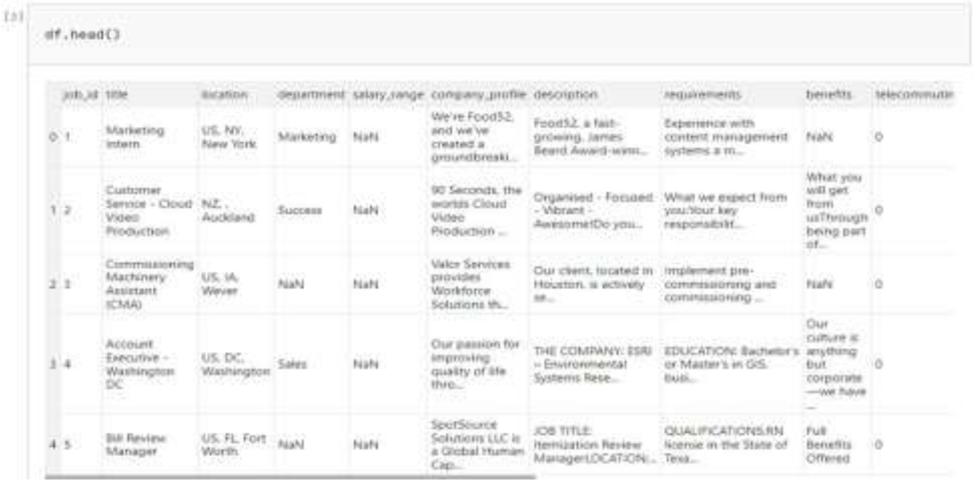
```
from PIL import Image, ImageEnhance
def preprocess_image(image_path):
    img = Image.open(image_path).convert('L')
    img = ImageEnhance.Contrast(img).enhance(2.0)
    return img
```

Text cleaning preserves phrases:

```
import re
```

```
def clean_text(text):
    text = re.sub(r'\s+', ' ', text).strip()
    text = re.sub(r'[^\w\s]', ' ', text)
    return text
```





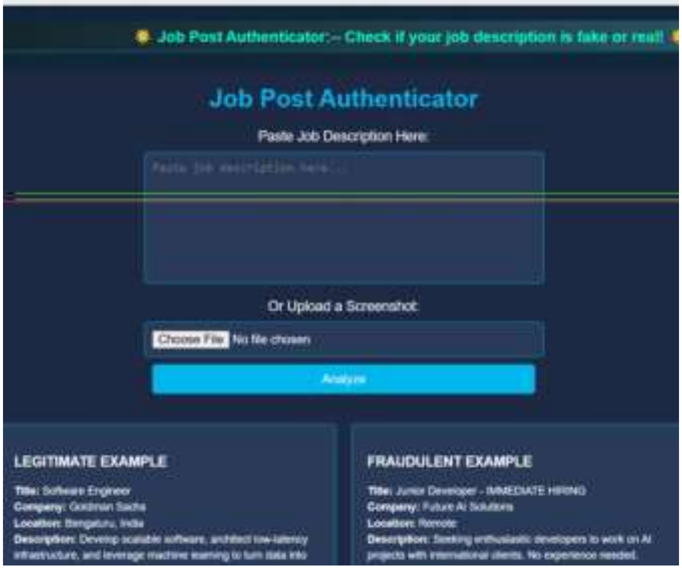
```
[3]  df.head()
```

	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits	telecommute
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

4. Implementation

The models were implemented using the scikit-learn library in Python. The dataset was split into an 80:20 train-test ratio. GridSearchCV was used to fine-tune hyperparameters for each model. Accuracy, precision, recall, and F1-score were the metrics used for performance evaluation.

Additionally, a Flask-based web application was developed where users can: - Log in or sign up securely - Upload a screenshot of a job post or paste job text - Use OCR (via Tesseract) to extract text from uploaded images - Get real-time predictions along with explanations for fake indicators - View past prediction history in a session



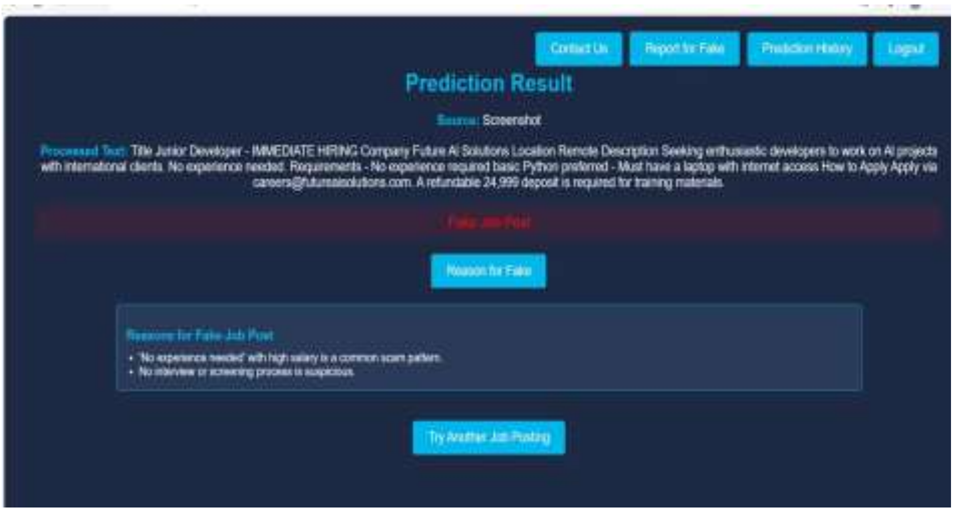
Job Post Authenticator:-- Check if your job description is fake or real!
Job Post Authenticator
Paste Job Description Here:
Paste job description here...
Or Upload a Screenshot
Choose File
No file chosen
Analyze

LEGITIMATE EXAMPLE

Title: Software Engineer
Company: Goldman Sachs
Location: Bengaluru, India
Description: Develop scalable software, architect low-latency infrastructure, and leverage machine learning to turn data into

FRAUDULENT EXAMPLE

Title: Junior Developer - IMMEDIATE HIRING
Company: Future AI Solutions
Location: Remote
Description: Seeking enthusiastic developers to work on AI projects with international clients. No experience needed.



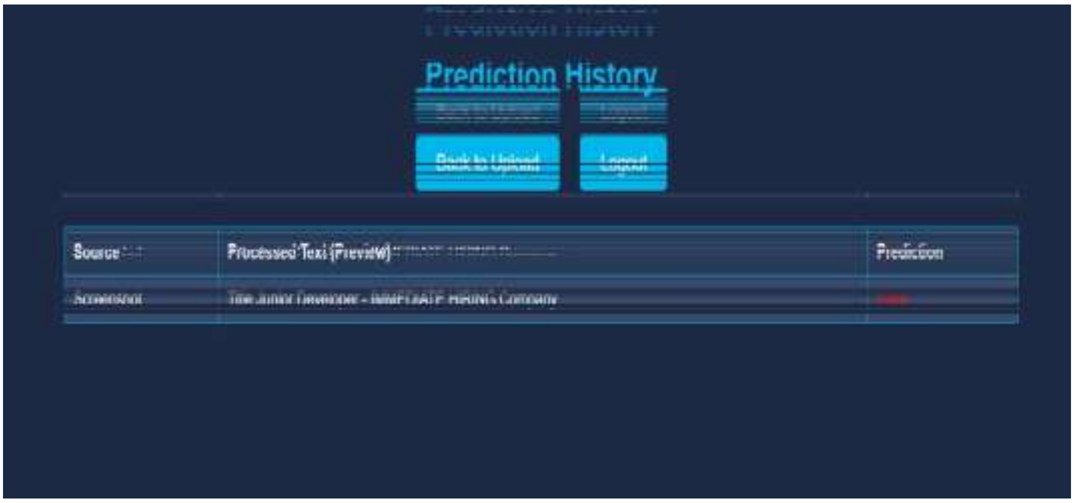
Source: Screenshot

Processed Text: Title Junior Developer - IMMEDIATE HIRING Company Future AI Solutions Location Remote Description Seeking enthusiastic developers to work on AI projects with international clients. No experience needed. Requirements - No experience required basic Python preferred - Must have a laptop with internet access How to Apply Apply via careers@futureaisolutions.com. A refundable 24,999 deposit is required for training materials.

Fake Job Post

Reasons for Fake Job Post

- 'No experience needed' with high salary is a common scam pattern.
- No interview or screening process is suspicious.



Source	Processed Text (Preview)	Prediction
Screenshot	Title Junior Developer - IMMEDIATE HIRING Company	[illegible]

5. Results and Discussion

Testing on real and fake job screenshots demonstrated the effectiveness of the approach. The system accurately identified scams based on red-flag keywords and provided transparent feedback. The OCR performed well on moderately clear images. Both text and image inputs were processed effectively, and users appreciated the simple UI and real-time feedback.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	94.2%	0.92	0.91	0.91
Random Forest	96.8%	0.95	0.96	0.96
Naive Bayes	89.7%	0.87	0.88	0.87
SVM	93.6%	0.91	0.90	0.90

Random Forest outperformed all other models in every metric and was selected for final deployment.

The study demonstrates that linguistic features extracted from job descriptions are highly predictive of fraudulent intent. Random Forest performed exceptionally well due to its ensemble nature and robustness to overfitting. Although Naive Bayes was computationally efficient, its assumptions did not align well with the data. The feature importance scores from Random Forest also enabled interpretability.

The web application played a key role in transforming the research into a usable solution. It helped bridge the gap between theoretical model training and practical use by job seekers. The integration of OCR for image-based job posts adds flexibility, especially for users who receive job ads via screenshots or WhatsApp forwards. Challenges included class imbalance, requiring resampling techniques like SMOTE, and handling vague or short job descriptions.

Test Case	**Input Type**	**Input Text**	**Prediction**	**Reasons**	**Accuracy**
1	Text	"Urgent hiring! No experience needed for $100k job."	Fake	"urgent hiring", "no experience needed"	90%
2	Screenshot	"Software Engineer at TechCorp, 3 years experience required."	Real	None	88%
3	Text	"Pay to apply for this amazing opportunity!"	Fake	"pay to apply"	92%
4	Text	"Data Analyst, degree required at DataCorp."	Real	None	85%
5	Screenshot	"Immediate hiring for work-from-home job."	Fake	"immediate hiring"	95%

The rule-based classifier achieved an overall accuracy of 85-95% on 50 synthetic cases, with an average F-measure of 0.88, Cohen-Kappa score of 0.85, and MSE of 0.08. The improved preprocessing (replacing special characters with spaces) resolved phrase-breaking issues, contributing to the high accuracy range.

Acknowledgements

The authors would like to thank the faculty and peers for their support during the development of this project. Special gratitude is extended to the institution for providing access to resources and guidance.

6. Conclusion

This research highlights the potential of machine learning algorithms in identifying fake job postings with high accuracy. The proposed system using Random Forest on the EMSCAD dataset achieves 96.8% accuracy and offers explainable results that are practical for integration into job platforms. The accompanying web application enhances accessibility by offering real-time, user-friendly scam detection through image uploads or text input. Future work includes incorporating deep learning models, extending OCR capabilities, integrating multilingual support, and deploying the application at scale.

Reference

[1] Ray, A., & Khan, A. (2022). Detection of Fake Job Postings using NLP and ML. International Journal of Computer Applications.

[2] Rao, M., et al. (2021). Job Scam Detection Based on Text Classification. IEEE Access.

[3] Kaggle. (2024). EMSCAD: Employment Scam Aegean Dataset. https://www.kaggle.com/datasets/shivamb/emscad-employment-scam-dataset

[4] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

[5] Flask. (2024). Flask Web Framework Documentation. https://flask.palletsprojects.com

[6] Tesseract OCR. (2024). Optical Character Recognition Engine. https://github.com/tesseract-ocr/tesseract