INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

EXPLAINABLE AI -JOB SCAM RISK SCORING SYSTEM

1st Pavithra Palanivel
Department of Advanced Computing and Analyics Vels Institute of Science, Technology& Advanced Studies
Chennai, India.
pavithrapalanivel2001@gmail.com

2 nd Dr. R. Mahalakshmi MCA, M.Phil., Ph D,
Associate Professor,
Department of Advanced Computing and Analytics Vels Institute of Science, Technology & Advanced Studies
Chennai, India rmahalakshmi.scs@vistas.ac.in

***Abstract*— With the emergence of online employment portals, the number of employment opportunities has also increased. However, this has also given rise to fraudulent employment opportunities. Detection of fraudulent employment opportunities through manual means is extremely difficult due to the sheer number of employment opportunities being published online. Hence, there is a need to implement intelligent systems to detect fraudulent employment opportunities in real time to avoid scams. The Explainable AI-Job Scam Risk Scoring System is a machine learning-based solution to automatically detect fraudulent employment opportunities. Natural Language Processing techniques have been implemented to process and analyze the content of the employment opportunity. Text data has been represented in numerical form using TF-IDF vectorization. This technique has been implemented to represent the importance of the words in the employment opportunity. To make the system transparent, SHAP-based Explainable AI has been implemented to identify the important features. The system is implemented in the form of a web application using Streamlit for real-time job scam detection. It processes the job description and makes predictions about whether a job posting is legitimate or a scam. Apart from prediction, the system also uses Explainable AI to increase transparency by explaining the factors that influenced the prediction outcome.**

Keywords*— *Explainable AI, Job Scam Detection, Machine Learning, NLP, TF-IDF, Logistic Regression, SHAP

INTRODUCTION

Online job platforms and professional networking sites have revolutionized the way employers and job seekers interact with one another. Organizations post job vacancies online, and they are able to reach potential candidates from different geographical locations. At the same time, job seekers are able to access a variety of job opportunities with minimal effort. This has revolutionized the hiring process and provided global employment opportunities. However, with this increase in online hiring, there are also opportunities for cyber criminals and fraudsters to carry out their activities. Fake job postings are one of the most common types of online fraud. Scammers post fake job postings that are legitimate and attractive to potential candidates. Fake job postings are one of the most common types of online fraud. Scammers post fake job advertisements that are legitimate and attractive to potential candidates. Job postings with high salaries, flexible working hours, remote jobs, or few qualifications are usually fake job postings. Scammers ask for personal information such as identification documents, bank account information, or payment for training, visas, or application fees once they get interested in the job posting. There are thousands of new job postings on online job platforms and it is very difficult to check all of them manually by humans. Furthermore, fake job postings are designed to resemble legitimate job postings, which makes it very difficult to detect them manually. Hence, there is a need to use Artificial Intelligence (AI) and Machine Learning (ML) to detect and prevent fake job postings.

LITERATURE SURVEY

Researchers have also explored various methods using machine learning, natural language processing, and data mining techniques for detecting and preventing such fraudulent activities. One of the initial methods for detecting fraudulent on-line content is the use of text classification methods. Researchers have used machine learning algorithms such as Naive Bayes, Support Vector Machine (SVM), and Logistic Regression for identifying patterns in the data. These algorithms use features such as keywords and sentence structures in the job description or advertisement for fraud detection. The studies revealed that machine learning-based classification methods can effectively identify patterns in the data for fraud detection applications. Though machine learning-based classification methods provide accurate results for fraud detection, most traditional AI systems are known for being a 'black box' approach, meaning that the system does not provide any explanation for its predictions. This has been a major challenge for users in trusting the system's predictions. In this regard, researchers have proposed a new approach called Explainable Artificial Intelligence (XAI).

PROPOSED WORK

The proposed system adopts a machine learning process that includes data collection, data preprocessing, feature extraction, training a model, prediction, and explanation. The system collects data from datasets that contain both fraudulent and legitimate job postings. The system preprocesses the data using various preprocessing techniques such as removing special characters, converting all the characters to lowercase, and removing stop words. The system uses TF-IDF vectorization to convert the job description into numerical features. The features represent the importance of each word in the dataset. The system uses a Logistic Regression classifier to train a model that learns

to distinguish fraudulent job postings from legitimate job postings. The system uses SHAP for explanation to determine the most important features in the system.

System Modules

Module 1: Data Processing Module: This module is responsible for the collection and preparation of the data that will be utilized for the training of the machine learning model. The data consists of job titles, company names, job descriptions, and legitimate and fraudulent labels. Preprocessing of the data is done by cleaning the data, removing punctuation, tokenization, and stopping words.

Module 2: Machine Learning Prediction Module: This module utilizes TF-IDF vectorization for converting the job description into numerical values. The machine learning model utilized here is Logistic Regression classification. Once the machine learning model is trained on the data, it predicts whether the job posting is legitimate or fraudulent and provides the confidence level.

Module 3: Explainable AI and Deployment Module: This module utilizes SHAP for explainability. The system highlights the words that were responsible for the prediction of the machine learning model. The deployment of the application utilizes Streamlit for the interface.

Architecture Diagram

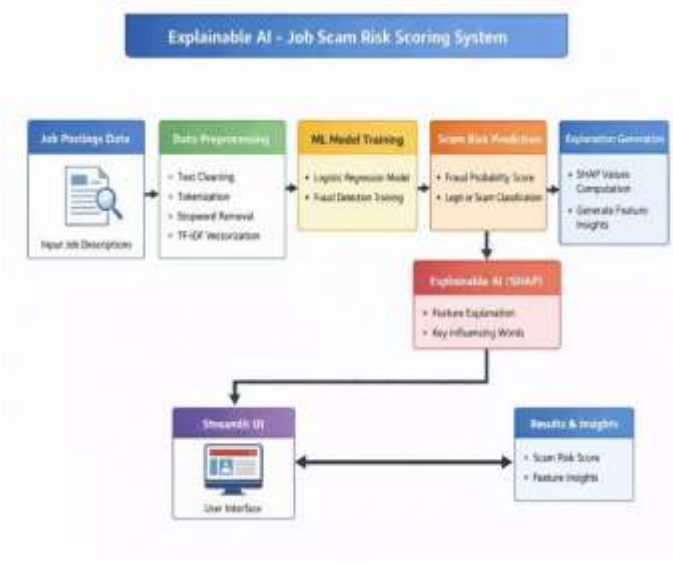


Figure 1: Architecture of Job scam risk scoring system

Project Phases / Methodologies

• Phase 1: Data Collection

Collect the job posting dataset from public repositories.

• Phase 2: Data Preprocessing

Clean and prepare the text data for analysis.

• Phase 3: Feature Engineering

Transform the data into TF-IDF vectors.

• Phase 4: Model Training

Train the Logistic Regression classifier.

• Phase 5: Explainability

Apply SHAP to interpret model predictions.

• Phase 6: Deployment

Deploy the model using Streamlit.

Pseudo Code

```
#MODEL

model = LogisticRegression( max_iter=1000, class_weight="balanced", solver="liblinear" )

pipeline = Pipeline( steps=[ ("preprocessor", preprocessor), ("classifier", model) ] )

#TRAIN MODEL
pipeline.fit(X_train, y_train)

#EVALUATION
y_pred = pipeline.predict(X_test)
y_proba = pipeline.predict_proba(X_test)[:, 1]
print("\nClassification Report:\n")
print(classification_report(y_test, y_pred))


roc_auc = roc_auc_score(y_test, y_proba)
print("ROC-AUC Score:", round(roc_auc, 4))

#SHAP EXPLAINER (FOR EXPLAINABLE AI)

# Extract trained components

tfidf = pipeline.named_steps["preprocessor"].named_transformers_["text"]
classifier = pipeline.named_steps["classifier"]

# Transform the full training dataset using the entire preprocessor
# This ensures the input to SHAP explainer matches the features the classifier was trained on (text + numeric)

X_train_transformed = pipeline.named_steps["preprocessor"].transform(X_train)
```

```
explainer = shap.LinearExplainer( classifier, X_train_transformed, feature_perturbation="interventional" )

#SAVE ARTIFACTS
pickle.dump(pipeline, open("model.pkl", "wb"))
pickle.dump(tfidf, open("vectorizer.pkl", "wb"))
pickle.dump(explainer, open("explainer.pkl", "wb"))
print("Model, Vectorizer, and SHAP Explainer saved successfully")
```

IMPLEMENTATION

The Explainable AI Job Scam Risk Scoring System can be implemented by utilizing the Python programming language and various libraries related to machine learning. The implementation steps are as follows:

1. **Programming Language:** Python
2. **Libraries Used:**
• Pandas– Handling data and data set processing
• NumPy– Numerical data processing
• Scikit-learn– Machine learning algorithms and model training
• SHAP– Explainable AI for model interpretation
• Streamlit– Web interface for user interaction
3. **Implementation Steps:**

Dataset Loading
The dataset containing job postings, both genuine and fraudulent, is loaded.
Data Preprocessing
Text data preprocessing for the job postings.
Feature Extraction
TF-IDF Vectorization for the job postings.
Model Training
Logistic Regression model training.
Model Evaluation
Accuracy, precision, recall, and F1-score are calculated.
Explainability
Integration of SHAP for determining the words that affect the model.
Model Deployment
Streamlit for creating a simple web interface for user input.

RESULTS AND DISCUSSION

The implemented system effectively analyzes the job postings and makes predictions regarding their legitimacy or fraudulent nature. The performance of the system can be considered satisfactory as the Logistic Regression model performs effectively in detecting fraudulent job postings. The TF-IDF feature extraction method performs effectively in detecting textual patterns in the job postings. The system also provides real-time scam risk scores for job postings. SHAP provides an explanation for the most influential words in detecting fraudulent job postings.

Example Output

Input Job Description:

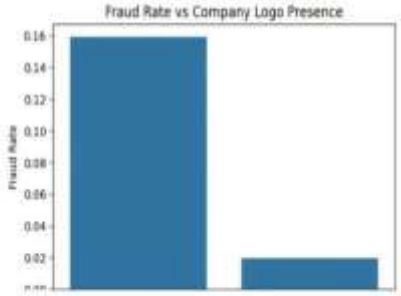
"Work from home opportunity with high salary. No experience required. Immediate joining with payment for training materials."

Output Prediction: Scam Risk Score: 0.87 (High Risk)
Prediction:
Potential Fraud Important Influencing Words:
"work from home"
"high salary"
"no experience required"

These keywords play a major role in helping the model classify the post as suspicious. Based on the results obtained, it is evident that machine learning algorithms can be used to identify patterns associated with fraudulent job postings. By analyzing the textual data, the system can identify suspicious keywords associated with fraudulent job postings. This is done by utilizing NLP techniques. Moreover, by incorporating TF-IDF vectorization, the system can assign weights to words in job postings. This helps the classification model differentiate between legitimate job postings and fraudulent job postings. Logistic Regression has been selected based on its simplicity, interpretability, and efficiency in text classification.

SCREEN SHOTS / CHARTS / GRAPHS

```
plt.figure(figsize=(6,4))
sns.barplot(
    data=logo_fraud,
    x="has_company_logo",
    y="fraudulent"
)
plt.ylabel("Fraud Rate")
plt.xlabel("")
plt.title("Fraud Rate vs Company Logo Presence")
plt.show()
```

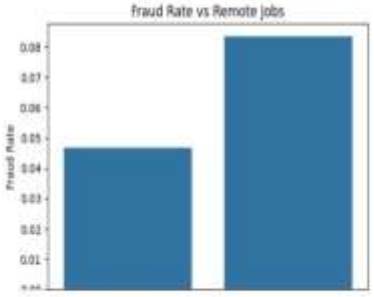



```
#Fraud rate wrt salary disclosure

import seaborn as sns
import matplotlib.pyplot as plt

salary_fraud = (
    job_fraud_df.groupby("salary_missing")["fraudulent"]
    .mean()
    .reset_index()
)

salary_fraud["salary_missing"] = salary_fraud["salary_missing"].map(
    {0: "Salary Provided", 1: "Salary Missing"}
)

plt.figure(figsize=(6,4))
sns.barplot(
    data=salary_fraud,
    x="salary_missing",
    y="fraudulent"
)
plt.ylabel("Fraud Rate")
plt.xlabel("")
plt.title("Fraud Rate vs Salary Disclosure")
plt.show()
```

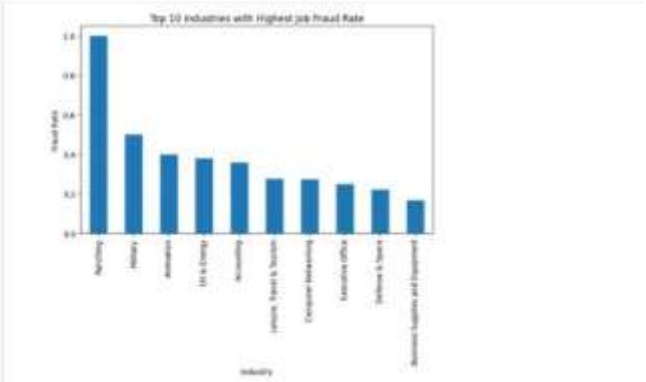

```
# Display top scam-indicating features
print("\n Top Features Increasing Scam Risk")
print(top_positive)

print("\n Top Features Indicating Genuine Jobs")
print(top_negative)
```

```
 Top Features Increasing Scam Risk
          feature  coefficient
2568         link     4.643486
321      aptitude     3.896084
2845        money     3.591341
2084  high school     3.586526
1             000     3.586162
1401         earn     3.296929
1123   data entry     3.260043
1516  engineering     3.121916
4067       signing     3.033113
3011          ohio     2.946114
4320        subsea     2.857633
1792    financing     2.839819
654          cash     2.739680
2529   leveraging     2.672569
3015       oil gas     2.646035

 Top Features Indicating Genuine Jobs
          feature  coefficient
4418         team    -3.038609
840     companies    -2.957775
1522      english    -2.539548
1994      growing    -2.272859
3934       search    -2.247788
452         based    -2.216621
69             50    -2.193721
3251        php     -2.121109
1287      digital    -2.057533
2697    marketing    -2.056489
3660   recruitment    -2.044394
1890          fun    -2.018663
4132     software    -1.964319
4820          web    -1.872804
1193     delivery    -1.842088
```

```
#Fraud rate vs telecommuting

remote_fraud = (
    job_fraud_df.groupby("telecommuting")["fraudulent"]
    .mean()
    .reset_index()
)

remote_fraud["telecommuting"] = remote_fraud["telecommuting"].map(
    {0: "On-site", 1: "Remote"}
)

plt.figure(figsize=(6,4))
sns.barplot(
    data=remote_fraud,
    x="telecommuting",
    y="fraudulent"
)
plt.ylabel("Fraud Rate")
plt.xlabel("")
plt.title("Fraud Rate vs Remote Jobs")
plt.show()
```



```
industry_fraud = (
    job_fraud_df.groupby("industry")["fraudulent"]
    .mean()
    .sort_values(ascending=False)
    .head(10)
)

plt.figure(figsize=(8,5))
industry_fraud.plot(kind="bar")
plt.ylabel("Fraud Rate")
plt.title("Top 10 Industries with Highest Job Fraud Rate")
plt.show()
```



CONCLUSION

The Explainable AI Job Scam Risk Scoring System shows the use of machine learning and natural language processing in identifying fraudulent job advertisements. The system works by processing the job descriptions provided in the advertisements. This is done using TF-IDF Vectorization, where the raw text data is transformed into numerical feature vectors that can be processed by machine learning algorithms. The Logistic Regression Classification Model is then applied to the processed data to identify patterns that distinguish between legitimate and fraudulent job advertisements. Based on this, the system identifies whether the advertisement is legitimate or fraudulent and provides a scam risk score. One of the significant contributions of this project is the inclusion of Explainable AI concepts using the SHAP (SHapley Additive Explanations) concept. However, there are several opportunities to enhance the proposed system. More data would allow the model to identify different types of job scams. Other advanced machine learning concepts like Deep Learning models can also be implemented to identify better contextual relationships in the data. Additionally, integration of the system with job portals may provide an opportunity to monitor fraudulent job postings in real time. In conclusion, the Explainable AI Job Scam Risk Scoring System provides a practical and intelligent solution to combat job scammers. By utilizing a combination of machine learning and explainable AI, the system not only helps identify fraudulent job postings but also provides an explanation for its predictions. This helps protect people from falling prey to scammers and creates a safer and more trustworthy environment.

REFERENCES

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[2] Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017.

[3] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *European Conference on Machine Learning*, 1998.

[4] Christopher D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[5] Trevor Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.

[6] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[7] Pandas Development Team, "Pandas: Powerful Python Data Analysis Toolkit," 2023.

[8] NumPy Developers, "NumPy Reference Guide," 2023.

[9] Streamlit Documentation, "Build Data Apps Faster," 2024.

[10] Logistic Regression model documentation in Scikit-learn official user guide.

[11] SHAP Documentation, "SHapley Additive Explanations for Machine Learning Models," 2024.

[12] TF-IDF documentation in Scikit-learn feature extraction module.

[13] Kaggle, "Fake Job Postings Dataset," used for fraud classification experiments.

[14] Natural Language Processing standard preprocessing methods from academic text classification literature.

[15] Explainable Artificial Intelligence research literature for transparent model interpretation.