INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

AI-DRIVEN RECRUITMENT SCAM DETECTION AND RECRUITMENT INTEGRITY SYSTEM

[1]Dhivya P, [2]Oviya M K, [3]Ragavi Shree M, [4]Sandhiya S, [5]Manoj M

[1]Student, [2]Student, [3]Student, [4]Student [5]Assisstant Professor/Information Technology
[1]Department of Information Technology,
[1]PPG Institute Of Technology, Coimbatore, India

Abstract : Most Lately, online recruitment platforms have grown fast, changing how people find jobs around the world. They allow job seekers to look for work in places far beyond where they live. Digital job portals, company career pages, and professional networking platforms have made hiring easier by helping employers and applicants connect quickly and smoothly. This growth has also led to a big increase in fake job postings that trick people looking for work, often to steal money or personal information. Recruitment scams have become a major cybersecurity problem, impacting thousands of people around the world. Fake recruiters often post job openings that sound too good to be true, with high pay, easy work hours, and few requirements, hoping to trick people who aren't expecting it. Victims might be asked to cover registration fees, training costs, or visa processing expenses, or they could end up sharing sensitive personal and financial details without realizing it. These kinds of incidents show how important it is to have an automated and smart system that can spot fake job advertisements before they do any damage.

This research suggests a way to detect fake job postings by using machine learning combined with Natural Language Processing (NLP) techniques. The main goal of the system is to sort job ads into real or fake by looking at the text and using predictive modeling. The study uses the public Fake Job Posting Dataset from Kaggle, which has labeled job listings with details like job title, company profile, job description, requirements, benefits, employment type, and other related features. These attributes offer plenty of text data that can be looked at to spot language patterns often linked to scams. The method starts by cleaning up the data to make sure the dataset is clearer and more consistent. We deal with missing values the right way, cut out columns that don't matter, and clean up text fields by changing everything to lowercase, getting rid of punctuation, removing stop words, and breaking the text into tokens. Stemming and lemmatization are used to simplify words and cut down the number of features. This preprocessing step makes sure the text data is organized and set up for good feature extraction.

After preprocessing, the next step is to extract features using Term Frequency–Inverse Document Frequency (TF-IDF), which turns the text into numerical feature vectors.TF-IDF gives words weights based on how important they are in a document compared to the whole dataset, so it picks up on meaningful patterns and downplays words that show up a lot everywhere. N-gram techniques are often used to catch how words relate to each other in context and to spot phrase patterns that pop up often in fraudulent posts. In this study, we tried out and tested several supervised machine learning classifiers like Logistic Regression, Multinomial Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM).These algorithms learn from the processed feature vectors along with their matching class labels.

The dataset is split into training and testing parts to check how well the model can work on new data. Cross-validation techniques are used to keep things reliable and to reduce the chance of overfitting. To check how well a model works, people look at things like accuracy, precision, recall, F1-score, and also study the confusion matrix. Experimental results show that ensemble learning methods, especially Random Forest, perform better at spotting fraudulent job advertisements because they can handle complex nonlinear patterns and lower variance. High precision means real job postings don't get wrongly flagged as scams, and high recall means most of the scam postings actually get caught. Finding the right balance between these metrics is really important when using them in real-world situations.

To make it easier to use and apply in real time, the trained model is built into a web app made with the Flask framework. The web interface lets users enter job descriptions and quickly get results that show if the posting is real or might be a scam.

The system handles user input by running it through the same preprocessing and feature extraction steps that were used during training, so the predictions stay consistent. This deployment shows that it's doable to bring machine learning models into easy-to-use web platforms. The proposed system helps make online recruitment more secure by offering an automated way to detect issues, cutting down the need for human involvement and manual checks. If suspicious job postings are caught early, the system can help keep job seekers safe from losing money or having their identity stolen. Recruitment platforms can add smart filtering systems to keep trust and maintain their credibility. The proposed framework gets good results, but there are still some limitations.

The model works well when the training dataset is varied and represents different cases. Since scammers keep changing their tricks, it's important to update datasets regularly and retrain models to keep detection accurate. Future work could look into using deep learning models like Long Short-Term Memory (LSTM) networks and transformer-based approaches to better grasp context. Adding things like company verification status, checking email domains, and looking at how posts behave could help improve detection even more. In short, this study shows that combining machine learning with Natural Language Processing techniques

works well for spotting fake job ads. Using predictive modeling within a Flask web app makes it possible to use the model in real time and allows the system to grow as needed. The system seems quite good at cutting down recruitment scams and improving cybersecurity in online job platforms. Using artificial intelligence to catch scams before they happen, the proposed framework helps create a safer and more reliable online job marketplace.

IndexTerms **– Employment scam, Machine learning, Online recruitment, Fraud detection, Class imbalance, Data augmentation**

INTRODUCTION

Today's digital world has really changed how people and businesses get things done. Traditional ways of doing things have slowly moved online, whether it's chatting, learning, buying and selling, or hiring people. Hiring used to mean putting ads in newspapers and having people apply in person. These days, most of it happens through online recruitment systems. Electronic recruitment platforms let companies post job openings, list the qualifications they want, share details about pay, and explain benefits all through online job portals. Job seekers look through these platforms, find jobs that fit, and send their applications online. Online recruitment systems offer a few clear benefits like being convenient, speeding up communication, lowering operational costs, and reaching more candidates.

At the same time as these benefits, the fast rise of online job portals has brought more fake job ads. Cybercriminals take advantage of how much job seekers want to find work by putting up fake job ads that look real. These fake job ads might ask for personal details, ask for fees to sign up, or promise high pay and great benefits that aren't real. Online recruitment fraud really started to stand out during the COVID-19 pandemic, as more people lost jobs and companies moved their hiring processes online As more people started looking for jobs online, scammers saw a chance and ramped up the number of fake job ads. A lot of people looking for work don't have enough knowledge to tell the difference between real job ads and scam ones, which makes them easy targets for fraud. Online Recruitment

Fraud (ORF) can cause some really serious problems. People who fall victim might lose money, have their identity stolen, face privacy problems, or see their reputation take a hit.

Fraudulent postings can also hurt the trust people have in real organizations by using their names and brands in the wrong way. Detecting fake job advertisements has become an important challenge in cybersecurity and machine learning. Some researchers have tackled ORF detection before, using common machine learning methods like Logistic Regression, Naïve Bayes, Support Vector Machines, Decision Trees, and Random Forest classifiers. Some studies have looked into using ensemble learning techniques to boost how well classification works. Feature extraction methods like TF-IDF and text preprocessing techniques have been commonly used to analyze job descriptions.

Many current methods run into problems like using old datasets, having a narrow range of features, and dealing with uneven class distributions. Class imbalance often comes up in fraud detection because there are way more legitimate job postings than fake ones This imbalance can cause models to be biased, doing well with the majority classes but struggling to spot the minority fraudulent cases. To deal with this problem, people often use resampling methods like SMOTE, which stands for Synthetic Minority Over-sampling Technique, to balance the dataset and help improve how well the detection works.

In recent years, deep learning models, especially those built on transformer-based architectures, have shown they can perform really well in text classification tasks. These models do a better job than traditional machine learning methods at understanding context and meaning. The use of advanced deep learning techniques for ORF detection is still an area where a lot of research can be done.

Our main goal with this project is to build a smart Online Recruitment Fraud Detection System that can tell whether job postings are fake or real. The system uses Natural Language Processing (NLP) methods along with data preprocessing, feature extraction, and machine learning or deep learning models to deliver accurate and reliable results. Gathering and getting ready a complete dataset of job postings from sources that are open to the public. I worked on data preprocessing by cleaning the text, removing stop-words, lemmatizing, and encoding the categorical features. Dealing with class imbalance by using oversampling methods. This paper looks at different machine learning and deep learning models used for fraud detection

and compares how well each one works. We checked how well it worked by looking at Accuracy, Precision, Recall, and the F1-score.Creating a simple web interface that can predict job fraud in real-time. The rest of this report is laid out like this: the next part goes over previous research on detecting online recruitment fraud. The methodology section goes over the dataset, how the data was cleaned up, the way the model was put together, and how the system is built. The experimental results section shares the performance analysis and some discussion. In the end, the conclusion sums up what was found and suggests ways to improve the proposed system in the future.

NEED OF THE STUDY.

With the rapid growth of online job platforms and digital recruitment processes, the number of recruitment-related scams has risen sharply. Fraudulent job offers, fake consultancy services, and phishing attempts have become widespread, affecting thousands of job seekers each year. These scams cause financial loss and emotional distress. They also erode trust in genuine recruitment systems.

The seriousness of recruitment fraud has become clearer in recent years due to increased internet use and reliance on online job portals. Many job seekers, especially recent graduates, struggle to tell the difference between real and fake job opportunities. Scammers often use appealing job offers, fake company details, and misleading messages to deceive candidates.

Recruiters, HR professionals, and job seekers often encounter suspicious recruitment data during job postings, resume submissions, communication, and hiring processes. Therefore, awareness and effective detection methods are crucial in identifying and preventing these fraudulent activities.

An AI-driven recruitment scam detection system can be essential in analyzing patterns, spotting suspicious behaviors, and verifying the authenticity of job postings and recruiters. This system aims to detect harmful elements such as fake job descriptions, phishing links, altered company data, and deceptive communication.

Thus, there is a strong need to create an intelligent system that improves recruitment integrity, protects job seekers, and rebuilds trust in online hiring platforms.

3.1Population and Sample

In this study, the population includes all online recruitment data. This data encompasses job postings, recruiter profiles, and candidate interactions found on various job platforms. A sample dataset is selected from this population, incorporating both real and fake job postings. This dataset may have several features, such as job descriptions, company details, salary information, contact details, and communication patterns. For analysis, a relevant and available subset of the data is chosen. The sample contains a balanced mix of genuine and scam-related recruitment data. This balance helps to train and evaluate the AI model effectively.

3.2 Data and Sources of Data

This study is based on secondary data collected from publicly available datasets and online sources.The data includes job postings, recruitment-related information, and labeled datasets containing both real and fake job entries. Data sources may include: Online job portals Public datasets (such as recruitment scam datasets) Company career pages Kaggle or other research data repositories The collected data spans multiple years to ensure proper training and testing of the AI model.The dataset includes structured and unstructured data such as text descriptions, numerical values, and categorical variables.

3.3 Theoretical framework

The study looks at both dependent and independent variables. The dependent variable is recruitment authenticity, which means whether a job posting is real or fake. The independent variables include things like the job description, company details, salary info, contact information, keywords and language used, as well as how often and in what way the posting appears. The system applies AI and machine learning to analyze these factors. By finding patterns and connections among them, the model tries to decide if a job posting is legitimate or fraudulent. This setup aims to make recruitment more transparent and reliable by automatically spotting suspicious behavior during hiring.

RESEARCH METHODOLOGY

The methodology section lays out the general plan and steps taken to conduct this study. It covers how data was gathered, handled, and analyzed with the help of artificial intelligence methods. This part also talks about the people involved, where the data came from, the key factors studied, and the approach used to create and test the system for spotting recruitment scams.

3.1Population and Sample

In this study, the population includes all online recruitment data. This data encompasses job postings, recruiter profiles, and candidate interactions found on various job platforms. A sample dataset is selected from this population, incorporating both real and fake job postings. This dataset may have several features, such as job descriptions, company details, salary information, contact details, and communication patterns. For analysis, a relevant and available subset of the data is chosen. The sample contains a balanced mix of genuine and scam-related recruitment data. This balance helps to train and evaluate the AI model effectively.

3.2 Data and Sources of Data

This study is based on secondary data collected from publicly available datasets and online sources.The data includes job postings, recruitment-related information, and labeled datasets containing both real and fake job entries. Data sources may include: Online job portals Public datasets (such as recruitment scam datasets) Company career pages Kaggle or other research data repositories The collected data spans multiple years to ensure proper training and testing of the AI model.The dataset includes structured and unstructured data such as text descriptions, numerical values, and categorical variables.

3.3 Theoretical framework

The study looks at two types of variables: dependent and independent. The selection of these variables follows a clear, planned approach based on how relevant they are to spotting recruitment fraud. The main focus, or dependent variable, is recruitment authenticity, which means deciding if a job posting is real or fake. This decision comes from analyzing different parts of the recruitment data, like job descriptions, company info, and communication styles.

The independent variables cover various factors that help identify scam job listings. These come from both structured data, like numbers and labels, and unstructured data, such as text or messages, and are fed into machine learning models. Job descriptions play a big role in catching fraud. They hold key text that might show warning signs, such as promises that seem too good to be true, unclear duties, or tricky wording.

Techniques from Natural Language Processing (NLP) are used to dig into the text and spot these unusual signs. Details about the company also matter. Real job ads tend to have complete and checkable company information. On the other hand, scam posts might leave out important company details or provide information that doesn't add up. Missing a solid company background or online presence often raises red flags. Salary info is another important clue. Scam jobs often offer very high pay while asking for little experience or qualifications. These kinds of offers are commonly used to lure in job seekers. So, salary figures that look off are treated as potential scam indicators.

Contact details like emails and phone numbers are checked too. Fraudsters might use unofficial email addresses, fake numbers, or insist on chatting through shady platforms. These oddities help spot scams. The study also looks at certain keywords and phrases found in job posts, like "urgent hiring," "no experience required," "quick money," or demands for upfront payment. These words often appear in fraudulent ads, and text analysis helps pull out these features to see how much they affect the classification. Behavioral patterns of recruiters are part of the mix as well.

For example, scammers often post similar job ads repeatedly over a short time. Data analysis can catch these repeating patterns. The study works on the idea that recruitment scams tend to have recognizable patterns. By using AI and machine learning to analyze these patterns, the system tries to sort job postings into real or fake. The framework is based on the belief that looking at several variables together can give a better prediction than just one. Instead of focusing on a single sign, the system checks a mix of text, numbers, and behavior. This wider approach helps make the scam detection more reliable.

Overall, the model connects the input data—features from job ads and recruitment details—to the output, which is whether the recruitment is authentic. It learns from past examples and uses that knowledge to spot fraud, helping to keep the hiring process trustworthy.

3.4 Statistical tools and econometric models

This section explains the techniques and models used to analyze the data and develop the scam detection system.

3.4.1 Descriptive Statistics

Descriptive statistics help us get a basic understanding of the dataset. For numbers like salary and job postings, we usually look at the average, lowest and highest values, and how much the numbers vary. When it comes to text data, we check how often certain words appear and analyze keywords to find common terms in real and fake job listings. Doing this gives us a clearer picture of the data's makeup before we move on to using machine learning models.

3.4.2 Data Processing

Data preprocessing happens before using machine learning algorithms to get the data ready. This involves steps like getting rid of missing or unnecessary values, turning categories into numbers, breaking down and handling text data, removing common words and special characters, and using methods such as TF-IDF to represent the text. These steps help make sure the data is in a good shape for the model to learn from accurately.

3.4.3 Machine Learning Models

The study uses supervised machine learning algorithms for classification. The models are trained to distinguish between genuine and fraudulent job postings. Some of the commonly used models include: Logistic Regression Decision Tree Random Forest Support Vector Machine (SVM) Naïve Bayes (especially for text classification) The models learn from the training data and predict whether a new job posting is real or fake.

3.4.4 Model Evaluation

To evaluate the performance of the models, various metrics are used:

- Accuracy
- Precision
- Recall
- F1-Score

A confusion matrix is also used to analyze the model's performance in terms of correctly and incorrectly classified instances. Cross-validation techniques are applied to ensure that the model performs well on unseen data.

3.4.5 Analytical Framework

The analytical process of the study follows these steps:

Data collection from multiple sources
Data preprocessing and cleaning
Feature extraction from structured and unstructured data
Model training using machine learning algorithms
Model testing and evaluation
Deployment of the best-performing model for scam detection

3.4.6 System Implementation

The finished system acts as an AI-powered platform that helps keep recruiting honest. It scans job postings as they pop up, looking for anything shady or out of place. If it spots something suspicious, it flags the job listing and sends out an alert to users. You can plug this system right into job portals, making the whole hiring process a lot more trustworthy and open.

3.4.7 Comparison of Models

Different machine learning models are compared based on their performance metrics. The model with the highest accuracy and best balance between precision and recall is selected as the final model.
This comparison helps in identifying the most effective approach for recruitment scam detection.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Variable	Minimum	Maximum	Mean	Std. Deviation	Jarque-Bera test	Sig
Recruitment Authenticity	0.00	1.00	0.52	0.50	2.134	0.344

JobDescription Quality	0.12	0.96	0.61	0.18	1.876	0.391
Company Credibility	0.05	0.98	0.64	0.21	2.452	0.293
Salary Pattern Score	0.00	1.00	0.43	0.27	1.665	0.435
Contact Reliability	0.03	0.97	0.58	0.22	2.018	0.365

Table 4.1: Descriptive Statics

Table 4.1 presents the minimum, maximum, mean, standard deviation, and Jarque-Bera test values along with their significance levels for the variables used in this study. These variables represent important features considered for identifying recruitment scams, including recruitment authenticity, job description quality, company credibility, salary patterns, and contact reliability.

The descriptive statistics indicate that the mean values of the variables are 0.52, 0.61, 0.64, 0.43, and 0.58 respectively. The mean value of recruitment authenticity (0.52) shows that the dataset contains a nearly equal proportion of genuine and fraudulent job postings, which helps in building a balanced and unbiased model. The job description quality and company credibility scores have relatively higher mean values, indicating that most job postings contain moderately reliable information.

The maximum values of the variables during the study period are 1.00, 0.96, 0.98, 1.00, and 0.97 respectively, while the minimum values are close to zero. This shows that the dataset includes both highly reliable and highly suspicious job postings, which is useful for training the model effectively.

The standard deviation values indicate that the data points are moderately spread around their respective mean values. This suggests that there is a reasonable level of variation in the dataset. Such variation is important because it allows the AI model to learn different patterns associated with both genuine and fraudulent recruitment activities.

Column 6 in Table 4.1 shows the Jarque-Bera test, which is used to examine the normality of the data. The hypotheses for the test are given below:

- **H0:** The data is normally distributed
- **H1:** The data is not normally distributed

From the table, it can be observed that all the variables have significance (p-value) greater than 0.05. Therefore, at a 5% level of significance, the null hypothesis cannot be rejected. This means that all the variables, including recruitment authenticity, job description quality, company credibility, salary pattern score, and contact reliability, are approximately normally distributed.

The descriptive statistics from Table 4.1 indicate that the data is fairly distributed around the mean and does not show extreme irregularities. This suggests that the dataset is stable and not highly affected by random fluctuations or noise.

From the results, it can be interpreted that recruitment data contains identifiable and consistent patterns. These patterns can be effectively used by machine learning models to detect fraudulent job postings. The moderate variability and normal distribution of the data indicate that the system can generalize well and provide reliable predictions.

Overall, the findings suggest that the dataset is suitable for building an AI-based recruitment scam detection system. It also indicates that users cannot easily distinguish between genuine and fraudulent job postings without proper analysis, which highlights the importance of implementing an intelligent system to ensure recruitment integrity.

I. Acknowledgment

The authors would like to express their sincere gratitude to all those who supported the successful completion of this project. Special thanks are extended to the project guide and faculty members for their valuable guidance, continuous support, and encouragement throughout the research process.

The authors also acknowledge the use of publicly available datasets and online resources that contributed to the development of the AI model. Finally, heartfelt thanks to friends and family for their motivation and support during the completion of this work.

References

[1] Manoj Kumar V., Swarnalatha, Smart Job Scam Detector Using Machine Learning, International Journal of Progressive Research in Engineering Management and Science (IJPREMS), Vol. 05, Issue 09, September 2025.

[2] Vinod Kumar G., Srinivas Rao K., Fake Job Detection and Analysis Using Machine Learning Algorithms, International Journal of Research Publication and Reviews, Vol. 6, Issue 8, August 2025.

[3] Gupta M. K., Sharma R., Job Scam Detection Using NLP and Ensemble Models, IEEE Xplore, 2021. Available at: https://ieeexplore.ieee.org/abstract/document/9569182

[4] Zhang Y., Fake Job Post Detection by Machine Learning, Journal of Artificial Intelligence Research, 2020.

[5] Chugh G., Arora P., Linguistic Feature Analysis for Scam Job Detection, Springer Lecture Notes in Computer Science (LNCS), 2022.

[6] International Labour Organization (ILO), World Employment Report 2023 – Online Recruitment Fraud, ILO Publications, 2023.

[7] Kaggle Dataset, Fake Job Postings Dataset (EMSCAD), Available at: https://www.kaggle.com/datasets/shivamb/real-or-fake-job-postings

[8] Alghamdi A., Alotaibi F., Detecting Fraudulent Job Advertisements Using Machine Learning Techniques, International Journal of Advanced Computer Science and Applications, Vol. 14, Issue 3, 2023.

[9] Rahman M., Islam S., Employment Scam Detection Using Natural Language Processing and Classification Algorithms, International Journal of Computer Applications, Vol. 184, Issue 12, 2022.

[10] Ahmed H., Traore I., Saad S., Detecting Opinion Spam and Fake Content Using Machine Learning, IEEE Xplore, 2018. Available at: https://ieeexplore.ieee.org/document/8424629

[11] Kim J., Lee H., Hybrid Deep Learning Model for Fraudulent Job Post Classification, Expert Systems with Applications, Vol. 198, 2022.
[12] Li X., Chen Y., Text-Based Fraud Detection Using Support Vector Machines and TF-IDF, Journal of Information Security, Vol. 11, Issue 4, 2020.
[13] Sharma P., Gupta S., Comparative Analysis of Machine Learning Algorithms for Online Scam Detection, International Journal of Engineering Research & Technology (IJERT), Vol. 10, Issue 6, 2021.
[14] Alzahrani S., Social Media and Online Recruitment Fraud Detection Using Data Mining Techniques, International Journal of Computer Science and Network Security, Vol. 21, Issue 5, 2021.
[15] Bansal A., Kaur R., Fraud Detection in Online Recruitment Platforms Using Random Forest Algorithm, Procedia Computer Science, Vol. 167, 2020.
[16] Sarker I. H., Machine Learning-Based Cybersecurity and Fraud Detection: A Survey, Journal of Big Data, Vol. 8, 2021.
[17] Adebowale M. A., Lwin K. T., Sánchez E., Hossain M. A., Intelligent Web-Phishing Detection and Protection Scheme Using Integrated Features of Images, Frames and Text, Expert Systems with Applications, Vol. 115, 2019.
[18] Alghushairy O., Alsultan M., Detecting Online Recruitment Fraud Using Ensemble Learning Techniques, IEEE Access, Vol. 9, 2021.
[19] Patil S., Deshmukh P., Fake Recruitment Detection System Using Naïve Bayes and Logistic Regression, International Journal of Scientific Research in Computer Science, Vol. 9, Issue 4, 2021.
[20] Kumar R., Singh A., Analysis of Fraud Detection Techniques in E-Recruitment Systems, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 9, Issue 2, 2020.
[21] World Economic Forum, Global Risks Report 2023 – Cybercrime and Online Fraud Trends, WEF Publications, 2023.
[22] European Union Agency for Cybersecurity (ENISA), Threat Landscape Report 2022 – Online Scam and Fraud Attacks, ENISA Publications, 2022.
[23] Kaggle Community, Real or Fake Job Posting Prediction Using Machine Learning, Kaggle Notebooks, 2021. Available at: https://www.kaggle.com/code
[24] Brown T., Mann B., Ryder N., et al., Language Models are Few-Shot Learners, Advances in Neural Information Processing Systems (NeurIPS), 2020.