

OfferCheck: AI-Based internship and Job Scam Detection

Mr. Adiki Vishwanath¹, Apoorva², Mahesh Billakanti³, Annavarapu Venkat⁴, Bupani Naveen⁵

¹Assistant Professor, Dept. of Computer Science and Engineering, HITAM, Telangana, India

²³⁴⁵Student, Dept. of Computer Science and Engineering, HITAM, Telangana, India

Abstract - The increased use of online job platforms and digital communication has created new opportunities for job seekers, but it has also increased the spread of fraudulent job and internship offers. Such offers may attract applicants with promises of high salaries, payment requirements, unofficial communication channels, urgent hiring messages, or guaranteed employment. Identifying these scams manually can be difficult and time-consuming, particularly for students and inexperienced job seekers. This paper presents OfferCheck, a web-based system for identifying potential risks in online job and internship offers. The system uses a machine learning classifier trained on TF-IDF representations of job descriptions for initial scam detection. It also applies predefined rules to identify suspicious characteristics, including payment requests, aggressive language, unrealistic income promises, unofficial communication channels, and employment guarantees. The system provides a prediction, confidence score, trust score, risk level, and reasons for suspicion. OfferCheck also supports text extraction from PDF and image files using PyMuPDF and EasyOCR, allowing uploaded documents to be analyzed. The application is implemented using React, Flask, and MySQL, with authentication and scan-history management. The system demonstrates how machine learning and rule-based analysis can support automated screening of online recruitment offers and provide users with understandable information about potentially fraudulent opportunities.

Key Words: OfferCheck, Job Scam Detection, Machine Learning, TF-IDF, Logistic Regression, Rule-Based Analysis, OCR, Online Recruitment.

1. INTRODUCTION

The rapid development of online recruitment platforms has changed the way students, graduates, and potential employees approach searching for internship and job opportunities. Job portals, social media, messaging applications, and communication channels has facilitated the process by enabling recruiters and candidates to find matches quickly than through traditional methods. However, the same technological developments also provide channels through which fraudulent recruiters can publish fraudulent job and internship offers. Online recruitment fraud became an important field of investigation as fraudulent postings can be designed to acquire funds or personal data or other sensitive information from the applicants. Previous research

identified online recruitment fraud as a growing field and explored using machine learning and natural language processing techniques for its detection.

Fraudulent recruitment offers can appear legitimate and may be difficult to identify at a glance. Fraudsters can use high salaries, joining requirements, guaranteed employment offers, registration or processing payments and unofficial communication channels to make these offers look realistic. Furthermore, the rising popularity of professionally designed and personalized fraudulent recruitment messages makes simple manual checking increasingly difficult. Recent research has therefore looked into the problem of finding automated approaches for detecting suspicious recruitment content using machine learning, deep learning, and explainable artificial intelligence techniques.

Several machine learning approaches have been proposed for online recruitment fraud detection. More recent studies explored transformer-based models and hybrid frameworks to capture deeper semantic patterns and offer more explainable predictions. Yet, many of the existing approaches focus primarily on the classification performance using datasets of job advertisements. A practical system for students and job applicants can additionally benefit from an explanation for the prediction, risk information, and the ability to analyze the job-offer content that was supplied in various formats.

To address this need, this paper presents OfferCheck: AI-Based internship and Job Scam Detection, a web-based system that helps the users to evaluate potentially fraudulent job and internship offers. The proposed system combines TF-IDF-based text representation, Logistic Regression classification, and rule-based analysis to detect suspicious recruitment content. Besides predicting if an offer is potentially fraudulent, the system generates a trust score, risk level, and reasons behind the prediction. The platform also combines Optical Character Recognition (OCR) to extract text from uploaded documents and images, thus allowing the users to analyze job offers that might not be available as directly-editable text.

The main contributions of the proposed system are as follows:

- Development of web-based platform for detecting potentially fraudulent job and internship offers.
- Combination of machine learning and rule-based analysis for recruitment scam detection.
- Generation of trust score and risk classification to make the prediction easier to understand.
- Identification of suspicious features that provide the users with reasons for the prediction.
- Use of OCR to analyze extracted text from uploaded job-offer documents and images.
- Integration of authentication and scan-history functionality to serve as a practical user-oriented detection platform.

2. LITERATURE REVIEW

Online recruitment has become an important part of the employment process, but the growth of digital recruitment platforms has also created opportunities for fraudulent job advertisements. Researchers have investigated online recruitment fraud using machine learning, natural language processing, text-mining techniques, and more recently deep-learning and explainable-AI methods. Existing studies provide the foundation for developing automated systems that can identify potentially fraudulent recruitment content.

2.1 Online Recruitment Fraud and Fake Job Advertisements

Early research investigated online recruitment fraud as a problem in automated detection and data mining. Vidros et al. studied the characteristics of fraudulent online recruitment and introduced the Employment Scam Aegean Dataset (EMSCAD), containing 17,880 annotated job advertisements. Their work highlighted the importance of automated approaches for identifying fraudulent recruitment content.

Subsequent research investigated different machine learning approaches for detecting fraudulent job advertisements. Naudé et al. examined different types of fraudulent jobs, including identity theft, corporate identity theft, and pyramid or multi-level marketing schemes. Their study evaluated lexical, syntactic, semantic, and contextual features using publicly available recruitment data. These studies demonstrate that the textual and structural characteristics of job advertisements can provide useful information for identifying suspicious recruitment activities.

2.2 Machine Learning Approaches for Recruitment Fraud Detection

Machine learning has been widely investigated for detecting fraudulent job advertisements. Another study developed a secured recruitment-fraud detection

framework, and other ensemble methods, have been evaluated for online recruitment fraud detection.

One study compared several machine learning algorithms for online recruitment fraud and reported strong performance from LightGBM and Gradient Boosting on its dataset. Another study developed a secured recruitment-fraud detection framework using AdaBoost, XGBoost, Voting, and Random Forest, reporting 98.374% accuracy for AdaBoost in its experimental setting. These studies indicate that conventional machine learning methods can provide effective baseline approaches for recruitment-fraud classification.

2.3 Text Representation and TF-IDF-Based Detection

Text representation plays an important role in detecting fraudulent job advertisements because job descriptions contain linguistic patterns that can help distinguish suspicious and legitimate offers. TF-IDF is commonly used to convert textual information into numerical features suitable for traditional machine learning algorithms.

Recent work has continued to investigate TF-IDF together with Logistic Regression for fake-job detection. A 2026 IEEE conference paper, for example, describes a fake-job detection system using TF-IDF text vectorization and Logistic Regression, with deployment through a Flask-based web interface. Similarly, comparative research on fraudulent recruitment advertisements has evaluated TF-IDF alongside traditional machine-learning algorithms and deep-learning approaches. Based on these findings, the proposed OfferCheck system uses TF-IDF to represent job-description text before applying Logistic Regression for classification.

2.4 Rule-Based Identification of Suspicious Job Characteristics

Machine-learning classification can identify patterns learned from training data, while explicit rules can identify clearly defined suspicious characteristics. Research on recruitment fraud has therefore also examined handcrafted features and rule-based indicators. Common suspicious characteristics reported in the literature include requests for payments, unusually attractive salary claims, suspicious communication methods, and misleading recruitment information. Earlier work compared machine-learning models using handcrafted features with text-based approaches and demonstrated the usefulness of manually identified characteristics for fraudulent-job detection.

In OfferCheck, predefined rules are used to identify indicators such as registration fees, processing fees, training fees, security deposits, WhatsApp or Telegram communication, urgent hiring language, guaranteed jobs, unrealistic income promises, and online-order processing

offers. These indicators are converted into understandable reasons that accompany the prediction.

2.5 Explainability and Risk Information in Fraud Detection

Recent research has increasingly focused not only on classifying job advertisements but also on explaining why a particular advertisement is considered suspicious. For example, XAI-Recruit uses deep learning together with LIME to provide token-level explanations for recruitment-fraud predictions. Another recent study developed an explainable recruitment-fraud framework that provides risk scores and term-level explanations to support human review.

These studies demonstrate the importance of providing understandable information alongside a classification result. OfferCheck follows a simpler implementation by providing a confidence score, trust score, risk level, and detected suspicious reasons.

2.6 OCR and Document-Based Job Offer Analysis

Job and internship offers are not always supplied as editable text. Applicants may receive offers as PDF documents or images, creating an additional requirement for extracting text before analysis. Recent research has also investigated OCR-based fake-job detection. One 2026 study developed a fake-job detection system incorporating EasyOCR, TF-IDF, machine learning, and an explanation component. OfferCheck addresses this practical requirement by using PyMuPDF (fitz) for extracting text from PDF files and EasyOCR for extracting text from image files. The extracted text is then processed using the same analysis pipeline as directly entered job descriptions.

2.7 Research Gap

Existing studies demonstrate the usefulness of machine learning, text representation, handcrafted features, and explainability for online recruitment fraud detection. However, many studies primarily evaluate classification models on datasets rather than presenting a user-oriented system that accepts job information in multiple formats and provides understandable risk information. Recent research has also explored transformer-based and more complex models. For example, transformer-based approaches using BERT and RoBERTa have been investigated for online recruitment fraud detection. The research gap addressed by OfferCheck is therefore the development of a relatively simple, practical web-based system that combines TF-IDF, Logistic Regression, predefined suspicious-feature rules, confidence score, trust scoring, risk classification, PDF text extraction, and image OCR within one application.

3. METHODOLOGY

The proposed system follows an experimental and application-oriented methodology for detecting potentially fraudulent job and internship offers. The system accepts job descriptions as text input and can also extract text from PDF and image files for analysis. The extracted or entered text is processed using TF-IDF and classified using a Logistic Regression model. In addition, predefined rules are used to identify suspicious characteristics such as payment requests, unrealistic salary promises, urgent hiring language, unofficial communication channels, and guaranteed employment claims. The methodology consists of data collection, text representation, machine learning model training, rule-based analysis, risk assessment, document text extraction, and system integration.

The overall workflow of the proposed system is as follows:

Job/Internship Offer → Text Input/File Upload → Text Extraction → TF-IDF Feature Extraction → Logistic Regression Classification → Rule-Based Analysis → Risk Assessment → Result Display → Scan History

3.1 Research Method

The proposed research follows an experimental approach because a machine learning model is trained and evaluated using a dataset containing genuine and scam job descriptions. The job-description text is converted into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF). These features are provided to a Logistic Regression classifier to determine whether the given job offer is potentially genuine or scam. The system additionally applies predefined rules to identify suspicious characteristics in the submitted offer. The results from the classification and rule-based analysis are used to provide understandable information to the user. The main objective of the methodology is to develop a practical system that can assist users in identifying potentially fraudulent job and internship offers.

3.2 Job Offer Dataset Collection

The dataset is used as the primary source for training the machine learning model. Each record contains information related to a job or internship offer, including the job description, company, job role, and corresponding label.

The dataset contains two classes:

- Genuine – legitimate job or internship offers.
- Scam – potentially fraudulent job or internship offers.

The dataset used for the experimental evaluation contains 1000 records. The dataset is divided into training and testing portions before model evaluation. The job description is used as the main input feature because fraudulent offers may contain specific textual patterns that can help distinguish them from genuine offers.

Table 1: Dataset Attributes

Attribute	Description
Job Description	Textual content of the job offer
Company	Name of the organization
Job Role	Position offered in the job advertisement
Label	Genuine or Scam

3.3 Text Feature Extraction Using TF-IDF

The textual job descriptions cannot be directly provided to the machine learning classifier. Therefore, the text is converted into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF assigns numerical importance to words based on their occurrence in a particular job description and across the dataset. This allows the machine learning model to identify useful textual patterns. In the proposed system, the TfidfVectorizer from Scikit-learn is used to convert job descriptions into a TF-IDF matrix. The resulting numerical representation is then supplied to the Logistic Regression model for classification.

3.4 Machine Learning Model

The proposed system uses Logistic Regression for classification of job and internship offers. Logistic Regression is a supervised machine learning classification algorithm that can be used for binary classification. In OfferCheck, the model learns patterns from labelled job descriptions and predicts whether a new job offer belongs to the genuine or scam category. The model is trained using the TF-IDF features generated from the dataset.

The training process is:

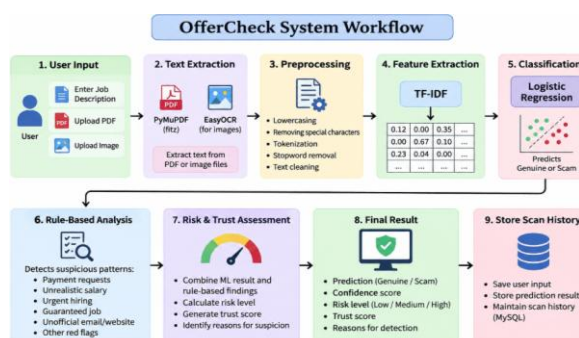


Fig. 1: OfferCheck System Workflow

Dataset → TF-IDF Vectorization → Training Data → Logistic Regression → Trained Model

The trained model and TF-IDF vectorizer are stored using the Joblib library. They can subsequently be loaded when a new job description is submitted for prediction.

3.5 Rule-Based Suspicious Feature Analysis

In addition to machine learning classification, OfferCheck uses predefined rules to identify suspicious characteristics in job and internship offers. The rules focus on indicators that may be associated with fraudulent recruitment, such as:

In the result list, you have:

- Requests for registration or processing fees.
- Requests for security deposits or other payments.
- Unrealistic salary or income promises.
- Urgent hiring statements.
- Guaranteed employment claims.
- Unofficial communication channels.
- Suspicious recruitment-related wording.

When one or more suspicious characteristics are detected, the system records the corresponding reason and uses it while generating the final risk information. This approach helps the user understand why an offer may be considered suspicious, rather than displaying only the classification result.

3.6 PDF and Image Text Extraction

Job and internship offers are not always available as editable text. Users may receive them as PDF documents or images. To support these inputs, OfferCheck provides text extraction functionality. PyMuPDF (fitz) is used to extract text from PDF files. For image files, EasyOCR is used to recognize and extract the text contained in the image. The extracted text is then passed to the same analysis process used for directly entered job descriptions. The workflow is:

PDF/Image → Text Extraction → Extracted Job Text → TF-IDF → Classification and Rule Analysis

This allows the system to analyze job offers received in different formats.

3.7 Risk and Result Generation

After classification and suspicious-feature analysis, OfferCheck presents the analysis results to the user. The result can include:

- Classification result

- Confidence score
- Trust Score
- Risk Level
- Reason for Suspicion

The suspicious indicators detected by the rule-based component are also presented to make the result easier to understand. The purpose of this stage is to provide users with more useful information than a simple Genuine/Scam prediction.

3.8 Web Application Implementation

OfferCheck is implemented as a web-based application. The frontend is developed using React, which provides the user interface for entering job descriptions and uploading files. The backend is implemented using Flask and Python and handles the machine learning prediction, text extraction, rule-based analysis, and API requests. MySQL is used to store user-related information and scan history. The main technologies used in the system are shown below.

Table 2: Technologies Used in OfferCheck

Component	Technology Used
Frontend	React
Backend	Flask, Python
Machine Learning	TF-IDF, Logistic Regression
OCR	EasyOCR
PDF Processing	PyMuPDF
Image Processing	Pillow
Database	MySQL
Authentication	bcrypt, PyJWT
API Testing	Postman
Model Storage	Joblib

3.9 Model Evaluation

The trained Logistic Regression model is evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score.

The experimental results obtained from the current model evaluation were:

Metric	Result
Accuracy	100.00%
Precision	100.00%
Recall	100.00%
F1-Score	100.00%

The confusion matrix obtained during the evaluation was:

	Predicted Genuine	Predicted Scam
Actual Genuine	100	0
Actual Scam	0	100

“The evaluation was performed on a test set containing 200 samples, including 100 genuine and 100 scam job descriptions.

4. RESULTS

The experimental results of the proposed OfferCheck system demonstrate its ability to analyze job and internship offers and identify potentially fraudulent recruitment content. The system uses TF-IDF-based text representation, Logistic Regression classification, rule-based analysis, and OCR-based text extraction.

4.1 Dataset Description

The dataset used in this study contains 1,000 job and internship offer records with two classes: Genuine and Scam. The dataset was synthetically generated using predefined templates representing common characteristics of genuine and potentially fraudulent recruitment offers. Each record contains the job description, company name, job role, and class label. Genuine records include characteristics such as official application procedures and the absence of registration fees, while scam records include suspicious characteristics such as payment requests, urgent hiring, guaranteed employment, and unofficial communication. The dataset was divided into 80% training data and 20% testing data, resulting in 800 training records and 200 testing records. The test set contained 100 genuine and 100 scam records.

Dataset details:

- Total number of records: 1,000
- Number of TF-IDF features: 113
- Classification classes: Genuine, Scam
- Text representation: TF-IDF
- Machine Learning model: Logistic Regression

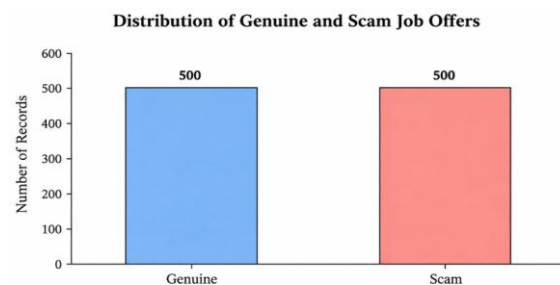


Fig. 2: Distribution of Genuine and Scam Job Offers
(Total Records: 1,000)

4.2 Feature Analysis

The job and internship descriptions were converted into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. The resulting TF-IDF matrix contained 1,000 records and 113 features. These features represent important terms and patterns present in the recruitment text and were provided as input to the Logistic Regression model for classification. In addition to the machine learning model, the system applies predefined rules to identify suspicious characteristics such as payment requests, unrealistic salary promises, aggressive language, unofficial communication channels, and guaranteed employment claims. These rule-based indicators provide additional information to support the classification result and explain the reasons for potential suspicion.

4.3 Logistic Regression Results

The Logistic Regression model was trained using the TF-IDF features to classify job and internship offers into genuine and scam categories. The trained model was successfully saved and used for prediction. On the evaluated test set of 200 job descriptions, the model achieved 100.00% accuracy, precision, recall, and F1-score. The confusion matrix showed that all 100 genuine and 100 scam offers were correctly classified, with no misclassified instances. For a given job description, the system produces a predicted class along with a confidence score.

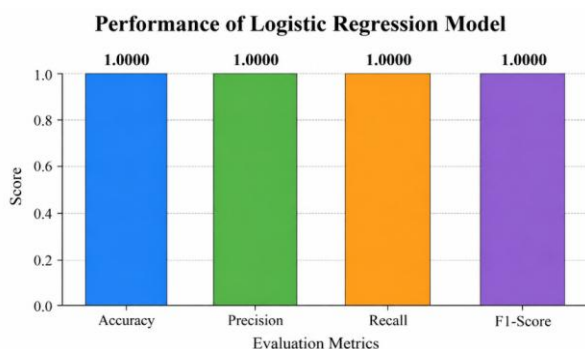


Fig. 3: Performance of the Logistic Regression Model (Accuracy, Precision, Recall and F1-Score)

5. DISCUSSION

The experimental results indicate that the proposed OfferCheck system can support the detection of potentially fraudulent job and internship opportunities. The system analyzes job-related information using Natural Language Processing, Machine Learning, rule-based analysis, and Optical Character Recognition (OCR). These techniques help identify suspicious characteristics present in job descriptions, offer letters, screenshots, and other recruitment-related content.

Features such as registration fee requests, unrealistic salary claims, urgent hiring messages, unofficial communication channels, and suspicious recruitment language provide useful information for identifying potential job scams. The TF-IDF representation converts textual job information into numerical features that can be processed by the Machine Learning model. The system also combines Machine Learning predictions with rule-based checks to provide additional information for classification and explain potential reasons for suspicion. OCR can extract text from uploaded images and PDF documents, allowing the extracted content to be analyzed by the detection system.

5.1 Performance Comparison and Analysis

The Logistic Regression model achieved an Accuracy of 100.00%, Precision of 100.00%, Recall of 100.00%, and F1-score of 100.00% on the evaluated test set. The test set contained 200 job descriptions, including 100 genuine and 100 scam samples. The confusion matrix showed that all 200 samples were correctly classified, with no misclassified instances.

These results indicate that the model was able to distinguish between the two classes on the evaluated dataset. However, the perfect score should be interpreted within the limitations of the dataset. The dataset contains synthetically generated job descriptions created using predefined templates. Therefore the results may not represent performance on diverse real-world recruitment advertisements. Evaluation using larger and independent real-world datasets would be required to further validate the model.

- **Accuracy:** 100.00%
- **Precision:** 100.00%
- **Recall:** 100.00%
- **F1-score:** 100.00%

5.2 Importance of Job-Related Features

Job descriptions contain textual patterns that can indicate potentially fraudulent recruitment activities. Suspicious characteristics such as requests for registration or processing fees, guaranteed employment, unrealistic earnings, urgent payment requests, unofficial communication channels, and suspicious recruitment language can contribute to the identification of potential scams. The TF-IDF technique converts job descriptions into numerical features based on the occurrence and importance of terms within the text. These features are provided to the Logistic Regression model for classification. In addition, rule-based checks provide further detection criteria based on predefined scam indicators.

5.3 Practical Significance of the Proposed System

The proposed OfferCheck system can assist students, job seekers, and recruitment platforms in identifying potentially fraudulent job and internship opportunities.

The system can contribute to:

- Early identification of suspicious job offers
- Reduction of recruitment scam risks
- Analysis of uploaded job documents and screenshots
- Identification of common scam indicators
- Support for safer job-search decisions
- Improved awareness of fraudulent recruitment practices

The proposed approach can be used for job offers received through websites, social media, email, messaging platforms, or other online recruitment channels.

6. CONCLUSION

Online recruitment scams have become a significant concern for job seekers because fraudulent offers can involve misleading job descriptions, fake recruitment processes, payment requests, and other suspicious activities. Identifying potentially fraudulent job opportunities before users interact with them can help reduce the risks associated with online recruitment. This study presented OfferCheck, a machine-learning-based job and internship scam detection system that combines TF-IDF, Logistic Regression, rule-based analysis, and OCR to analyze recruitment-related information. The system accepts job descriptions and recruitment-related documents or images and analyzes them for characteristics associated with potentially fraudulent offers.

The Logistic Regression model achieved 100.00% Accuracy, Precision, Recall, and F1-score on the evaluated test dataset. However, these results are limited to the current synthetically generated dataset and should not be interpreted as guaranteed performance on real-world job advertisements. Further evaluation using larger and independent real-world datasets is required. The developed system focuses on identifying potentially fraudulent job opportunities rather than guaranteeing that a particular company or job offer is legitimate. The prediction results and identified suspicious characteristics can help users make more informed decisions before responding to recruitment offers.

Overall, the study demonstrates the potential of Machine Learning, TF-IDF-based text analysis, rule-based analysis, and OCR as an approach for supporting the detection of

online recruitment scams and improving user awareness and safety.

6.2 FUTURE SCOPE

The proposed work can be extended in several ways. OfferCheck can be enhanced by using larger real-world datasets and advanced NLP models to improve its ability to identify diverse recruitment scams. The system can support real-time company and website verification, analyze emails and social-media job offers, and detect newly emerging scam patterns. It can also be extended with multilingual support, mobile application integration, and continuous model updates to provide comprehensive assistance for job seekers.

REFERENCES

- 1) S. Vidros, C. Koliass, G. Kambourakis, and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset," *Future Internet*, vol. 9, no. 1, p. 6, 2017. doi: 10.3390/fi9010006.
- 2) R. Wang et al., "A Comparative Analysis of Fraudulent Recruitment Advertisement Detection Methods in the IoT Environment," *Journal of Sensors*, 2022.
- 3) R. Frissen, K. J. Adebayo, and R. Nanda, "A machine learning approach to recognize bias and discrimination in job advertisements," *AI & Society*, vol. 38, pp. 1025–1038, 2023. doi: 10.1007/s00146-022-01574-0.
- 4) M. Naudé, K. J. Adebayo, and R. Nanda, "A Machine Learning Approach to Detecting Fraudulent Job Types," *AI & Society*, vol. 38, no. 2, pp. 1013–1024, 2023, doi: 10.1007/s00146-022-01469-0.
- 5) K. Taneja, J. Vashishtha, and S. Ratnoo, "Fraud-BERT: Transformer Based Context Aware Online Recruitment Fraud Detection," *Discover Computing*, 2025, doi: 10.1007/s10791-025-09502-8.
- 6) "A Hybrid Approach to Fake Job Detection Using NLP and Machine Learning," 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG). doi: 10.1109/ICTBIG68706.2025.11323989.
- 7) "Information Engineering for Fake Job Postings Classification in Electronic Business Based on Machine Learning Technology," *International Journal of Information Engineering and Electronic Business*, 2025. doi: 10.5815/ijieeb.2025.05.07.
- 8) "Career Verify – A Fake Job Posting Detection," *IEEE*, 2026. doi: 10.1109/AIDE69088.2026.11544522.
- 9) R. A. Jobu, A. Baral, S. M. S. Rahman, and M. N. Probor, "An explainable, cost-aware hybrid framework for

fraudulent online job posting detection,” Array, vol. 31, 101191, 2026. doi: 10.1016/j.array.2026.101191.

- 10) A. M. T. Hermoso, B. B. O. Lastimosa, L. S. Jardeleza, and J. A. Bacus, “‘Job or Jab?’: Detecting Fake Job Postings Using NLP,” Proceedings of the 2025 3rd International Conference on Artificial Intelligence and Power Engineering (AIPE2025), pp.14–19. doi10.1109/AIPE67885.2025.00010.