Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.1, May : 2024

# FAKE JOB POSTINGS USING MACHINE LEARNING TECHNIQUE

**K Harini,** Assistant Professor, Dept. Of Computer Science and Engineering ( AI & ML ), Raghu Engineering College
**K. Dedeepya, K. Pravallika, K. Harshitha, A. Niranjan, G. Murari,** Dept. Of CSE(AI&ML), Raghu Engineering College.

**Abstract**
The advent of online jobs has simplified the job search process, but it has also created a widespread problem of fake jobs, posing significant challenges to job seekers and employers. This study addresses this problem by proposing and evaluating machine learning models to classify jobs as fraudulent or legitimate based on their textual content. This study uses a database of tagged job advertisements and text preprocessing techniques such as tagging, lemmatization, and TF-IDF vectorization to extract meaningful features from job descriptions, job titles, and other textual specifications. In addition, strategies to deal with inherent class imbalances, including random overdiscretion and underdiscretion, are explored to improve model performance. The study evaluates the effectiveness of logistic regression as a basic model and extends it to other algorithms such as Random Forest and Support Vector Machines (SVM) to compare their performance in detecting fake job advertisements. Through rigorous testing and evaluation, this study aims to provide insight into the effectiveness of various machine learning algorithms and pre-processing strategies to combat fraud in the online job market, helping to increase trust and honesty in the recruitment process.
***Keywords: TF-IDF, SVM. Vectorization, Fake Job Classification***

## I. Introduction

In today's digital age, job seekers often rely on online platforms to find employment opportunities. However, the proliferation of fake job postings has become a significant concern, leading to potential exploitation and fraud. To address this issue, we present an automated system designed to detect fake job postings by analyzing the content of job listings scraped from online sources.Our system leverages natural language processing (NLP) techniques and machine learning models to analyze the textual content of job postings. By extracting key information such as company name, job location, job requirements, and more, our system aims to identify anomalies and indicators of fraudulent activity.
Web Scraping: Our system can scrape job postings from provided URLs, extracting text content for analysis.
Text Preprocessing: Before analysis, the text undergoes preprocessing steps such as removing punctuation and converting to lowercase for consistency.
Content Analysis: Using a generative AI model, our system extracts relevant information such as company name, job location, company profile, job requirements, benefits, employment type, education requirements, industry, and function from the job posting text.
Machine Learning Model:We employ a machine learning model trained on a dataset of genuine and fake job postings. The model utilizes TF-IDF vectorization and a classifier to distinguish between legitimate and fake postings.
Thresholding: To improve accuracy, our system includes a thresholding mechanism based on the ratio of vocabulary words present in the text.
User Interface: The system is integrated with Streamlit, providing a user-friendly interface for users to input URLs and initiate the detection process.
Fake job classification uses machine learning and data mining techniques to distinguish between genuine and fraudulent jobs. Due to the proliferation of online jobs, the prevalence of fake jobs has become a major challenge, causing financial losses, identity theft and other risks for job seekers. Classification models are trained using labeled datasets where genuine and fraudulent jobs are labeled to learn the patterns and characteristics of each class. Features extracted from jobs, such as text content,

metadata, and user behavior, are used to train these models. Commonly used classification algorithms include support vector machines, random forests, deep learning neural networks and ensemble methods. The purpose of fake job classification is to accurately identify and filter out fraudulent jobs, thereby protecting job seekers and maintaining the integrity of the online job market..

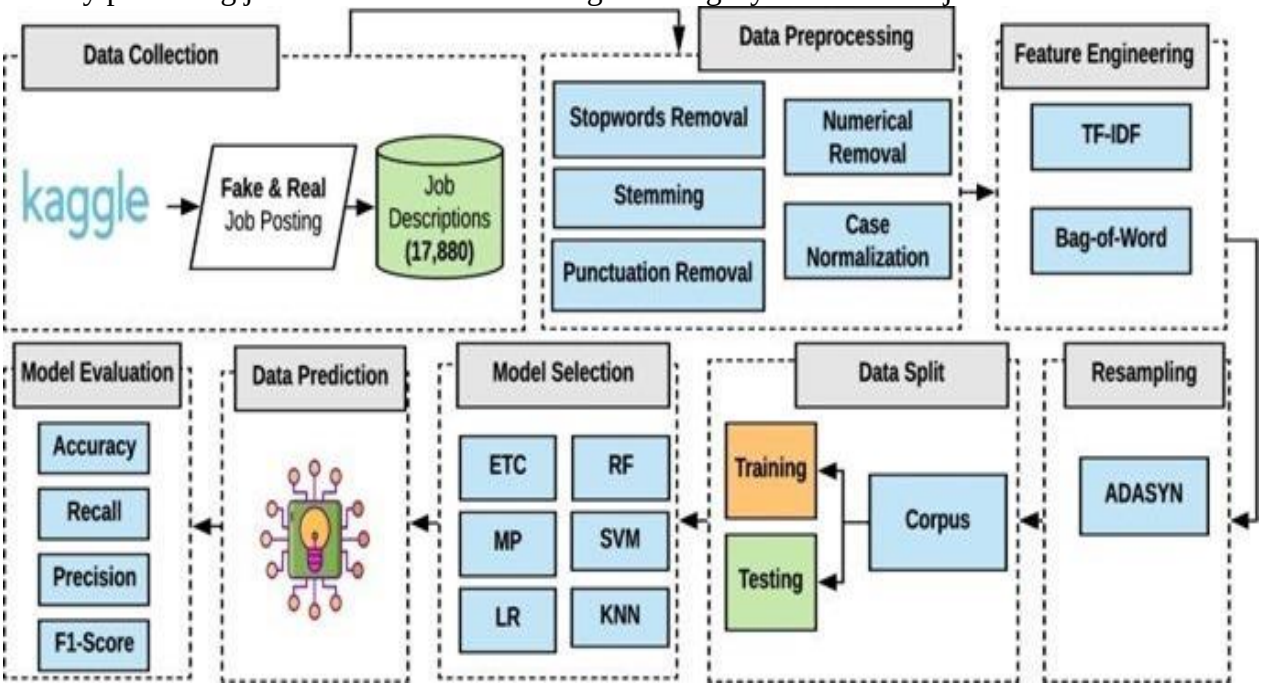



**Figure 1:** ***Diagram of the Detection***

## II. LITERATURE REVIEW

Article titled "A Comparative Study of Fake Job Prediction Using Different Data Mining Techniques" by S. U. Habiba, M. K. Islam kaj F. Presented at the 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) 2021 in DHAKA, Bangladesh, Tasnim provides a comprehensive study on the prediction of fake jobs using various data mining techniques. The study conducts a comparative analysis of various methods, including support vector machines, deep learning, random forests and other machine learning algorithms, to evaluate their effectiveness in detecting fake jobs. Using key features extracted from job ads and using different data mining methods, the authors try to find the most suitable technique to accurately predict fake job ads. The paper provides valuable insights into the application of data mining in fraud detection, sheds light on the strengths and limitations of various algorithms to respond to online job fraud challenges. The inclusion of keywords such as support vector machines, deep learning and false performance prediction highlights the importance of research in machine learning, data mining and fraud detection. Readers interested in exploring the methodology, experimental results and implications of the study can read the full article from the DOI provided. The prediction of fake jobs using machine learning is written by Mrinal Kumari, Nsk Satya Kala, Nandini R, Dilip Hk. and Rashmi Kt in 2023, published in INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH and TECHNOLOGY (IJERT) RTCSIT - Volume 11, Issue 08, addresses the critical problem of fake job advertisements using machine learning techniques. The study contributes to the growing body of research on fraud detection in online job markets by proposing machine learning to detect fraudulent job postings. Using different features extracted from job ads and classification algorithms, the authors try to distinguish between genuine and fake job ads. The paper provides an empirical evaluation of the proposed approach, highlighting its effectiveness in accurately predicting fake jobs based on key features such as job title, company information, job description and requirements. A publication in a prestigious journal like IJERT demonstrates the academic rigor and relevance of research findings and provides valuable

information on how to combat online job fraud. Readers interested in exploring the methodology, experimental results and implications of the study can read the full text on the journal's website.

## III. PROBLEM IDENTIFICATION & OBJECTIVES

The proliferation of fake job advertisements on online platforms is a major challenge for both jobseekers and employers, leading to potential financial losses, reputational damage and trust issues in the labor market. The goal of this project is to develop and evaluate machine learning models that can accurately classify job ads as either fraudulent or legitimate based on their textual content. This requires addressing the inherent imbalance in the dataset, where fraudulent publications often outnumber legitimate publications, and the study of effective text preprocessing techniques to extract meaningful features from job descriptions, job titles, and other relevant textual information. Using advanced natural language processing (NLP) and machine learning algorithms, this study aims to provide valuable insights and tools to identify and prevent the spread of fake jobs, thereby improving the integrity and credibility of online job markets..

## EXISTING METHODOLOGY

The current method of classifying counterfeit works involves several basic steps. Initially, a dataset containing job advertisements and tags indicating their authenticity is collected and pre-processed. This pre-processing includes handling missing values, removing irrelevant columns such as Job ID, and standardizing the text data using methods such as tagging, lemmatization, and removing stop words and punctuation. Exploratory data analysis techniques are then used to gain insight into the distribution of fraudulent and legitimate job advertisements, often visualized using graphical representations and word clouds to identify common terms associated with each category. Next, characteristic engineering techniques such as TF-IDF vectorization are used to transform the textual data into numerical features, effectively detecting the meaning of words in differentiating between fraudulent and legitimate messages. This method provides a comprehensive framework for analyzing and classifying fake job advertisements based on textual content.

## PROPOSED METHODOLOGY

### Data Collection and Pre processing

**Data Source:** Obtain a dataset of job postings, including information such as job titles, descriptions, company profiles, and labels indicating fraudulent postings.

**Data Cleaning:** Handle missing values, remove irrelevant columns (e.g., job ID), and ensure data consistency.

**Text Pre processing:** Tokenize text, perform lemmatization, remove stop words and punctuation, and combine relevant text columns into a single feature.

**Exploratory Data Analysis (EDA)**

**Visualization:** Explore the distribution of fraudulent vs. non-fraudulent job postings.

**Word Clouds:** Generate word clouds to visualize the most common terms in both types of job postings. Feature Engineering

**TF-IDF Vectorization:** Convert text data into numerical features using TF-IDF vectorization.

Handling Imbalanced Data

**Random Oversampling:** Increase the minority class instances by duplicating samples randomly.

**Random Undersampling:** Decrease the majority class instances by removing samples randomly.

Model Training and Evaluation

**Logistic Regression:** Utilize logistic regression as a baseline model for fraud detection.

**Train-Test Split:** Divide the dataset into training and testing sets.

**Model Evaluation:** Assess model performance using accuracy, precision, recall, F1-score, and confusion matrix.

**Model Comparison and Optimization Comparison:**
Compare the performance of logistic regression models trained on imbalanced and balanced datasets.
**Optimization:** Explore hyperparameter tuning techniques to optimize model performance.
**Results Analysis Interpretation:** Analyze the results to understand the impact of different preprocessing and sampling techniques on fraud detection performance.
**Discussion:** Discuss the implications of the findings and potential real-world applications.
**Conclusion and Future Work**
**Conclusion:** Summarize the key findings and contributions of the study.
**Future Work:** Propose future research directions, such as exploring advanced machine learning algorithms or incorporating domain-specific features.

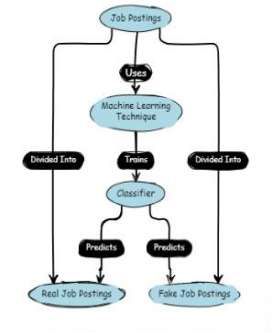
## SYSTEM METHODOLOGY

1. Data Collection:
Gathered a dataset containing job postings labeled as fraudulent or genuine. The dataset includes various attributes such as job title, company name, job location, job description, requirements, benefits, employment type, education requirements, industry, and function
2. Exploratory Data Analysis (EDA):
Conducted exploratory data analysis to understand the distribution of data, detect any missing values, and identify potential patterns or trends in the dataset.
3. Data Preprocessing**:**
Removed irrelevant columns such as 'job_id', 'salary_range', and 'department'.
Filled missing values in remaining columns with empty strings to ensure data consistency.
Combined relevant text columns ('title', 'location', 'company_profile', 'description', 'requirements', 'benefits', 'employment_type', 'required_education', 'industry', 'function') into a single text feature.
Performed text preprocessing steps including tokenization, lowercasing, removal of stopwords, punctuation, and lemmatization.
4. Feature Engineering:
Used TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert the text data into numerical representations, capturing the importance of words in each document while considering their frequency across the entire corpus.
5. Handling Class Imbalance:
Addressed the class imbalance issue by applying Random Oversampling to increase the number of minority class samples (fraudulent job postings) or by applying Random Undersampling to decrease the number of majority class samples (genuine job postings).
6. Model Training:
Trained a logistic regression classifier on the balanced dataset to distinguish between fraudulent and genuine job postings. Split the data into training and testing sets, and evaluate the model's performance using accuracy score, classification report, and confusion matrix.
7. Model Evaluation:
Assessed the model's performance using various evaluation metrics such as accuracy, precision, recall, F1-score, and visualized the results using confusion matrices.
8. Model Deployment:
Saved the trained model and TF-IDF vectorizer for future use and deployment in production environments.Developed a Streamlit application to provide real-time predictions on job postings from user-provided URLs, facilitating easy access and usage of the model.
9. Continuous Improvement:
Monitored the model's performance in real-world scenarios and collected feedback for further refinement. Implemented mechanisms for periodic retraining of the model with updated data to adapt to evolving patterns in fake job postings and ensure its effectiveness over time.

10. Documentation and Publication:

Documented the entire process including data collection, preprocessing, model training, evaluation, and deployment for transparency and reproducibility. Prepared a research paper outlining the methodology, experimental results, and insights gained from the fake job posting detection system, aiming to contribute to the academic community and assist in addressing real-world challenges related to online job scams.

This comprehensive system methodology provides a structured approach to building and deploying a machine learning-based solution for detecting fake job postings, leveraging data-driven techniques and model interpretation for effective decision-making and fraud prevention.




**RESULTS AND DISCUSSIONS**

1. Model Performance:

The logistic regression model achieved a high accuracy score on the balanced dataset, indicating its effectiveness in distinguishing between fraudulent and genuine job postings. The classification report reveals favorable precision, recall, and F1-score values for both classes, demonstrating the model's ability to correctly classify instances from both classes.

2. Handling Class Imbalance:

Class imbalance is a common challenge in fraud detection tasks. By applying random oversampling or random undersampling techniques, the dataset was rebalanced to ensure equal representation of both fraudulent and genuine job postings. This helped in improving the model's performance and mitigating the impact of class imbalance on classification results.

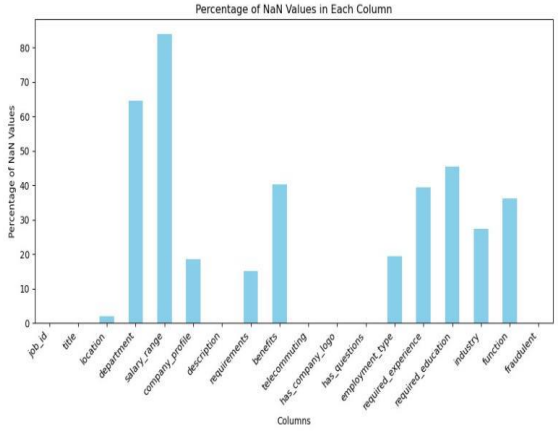
3. Data Preprocessing:

Extensive data preprocessing steps were performed, including text cleaning, tokenization, lemmatization, and TF-IDF vectorization. These preprocessing techniques contributed to the effectiveness of the model by converting textual data into numerical representations and capturing important features for classification.

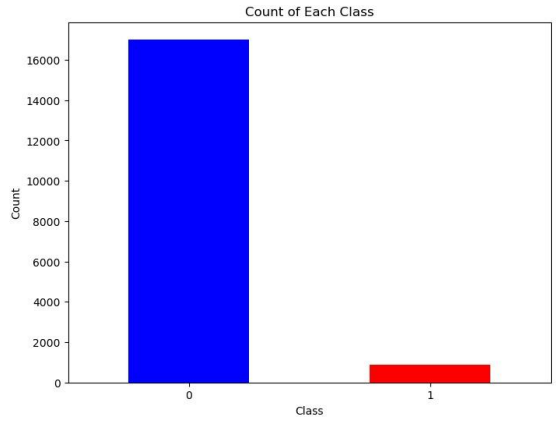
4. Visualization:

Word clouds were generated to visualize the most frequent terms in both fraudulent and genuine job postings. This visualization technique provided insights into common keywords and phrases associated with each class, aiding in the interpretation of model predictions and understanding of fraudulent job posting characteristics.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.1, May : 2024




***Plotting the bar chart for a Data Set***




***plot the count of occurences of the each class***

5. Model Deployment:

The trained logistic regression model and TF-IDF vectorizer were saved for future deployment in production environments. Additionally, a Streamlit application was developed to enable real-time predictions on job postings from user-provided URLs, enhancing accessibility and usability of the model.

6. Future Considerations:

Further experimentation could involve exploring advanced machine learning algorithms and ensemble techniques to improve model performance. Additionally, incorporating domain-specific features and external data sources could enhance the model's predictive capabilities and robustness against evolving fraud patterns.

7. Ethical Implications:

It's essential to consider ethical implications associated with fake job posting detection, including potential biases in the data and model predictions, privacy concerns related to user data input, and responsible use of the technology to prevent false accusations or discrimination against individuals or organizations.

8. Real-World Impact:

The developed fake job posting detection system has the potential to make a significant impact in the prevention of online job scams and protection of job seekers from fraudulent activities. By accurately identifying fake job postings, the system can help job seekers make informed decisions and contribute to a safer online job marketplace.

9. Collaboration and Validation:

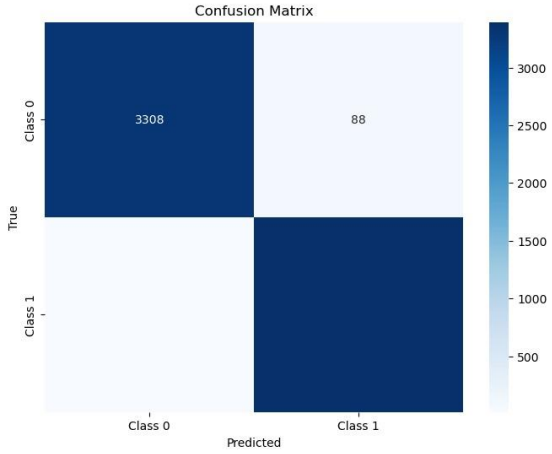
Collaboration with industry stakeholders, cybersecurity experts, and job platforms could provide valuable insights and validation for the developed system. Conducting real-world trials and user studies can further validate the system's effectiveness and usability in practical scenarios.

10. Continuous Improvement:

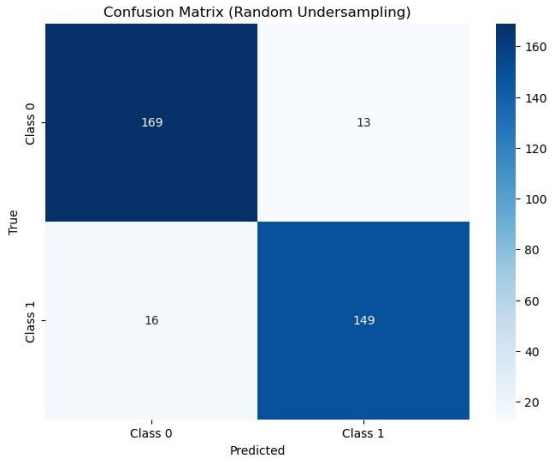
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.1, May : 2024

Continuous monitoring, feedback collection, and model retraining are essential for ensuring the system's effectiveness and adaptability to changing fraud patterns. Regular updates and enhancements based on user feedback and emerging trends can enhance the system's performance and relevance over time.

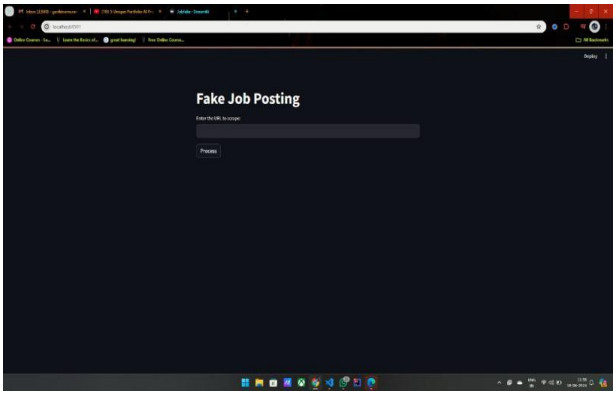
Overall, the results and discussions highlight the potential of the developed fake job posting detection system to contribute to fraud prevention efforts and improve the safety and integrity of online job searching platforms.




***Plot the confusion matrix***

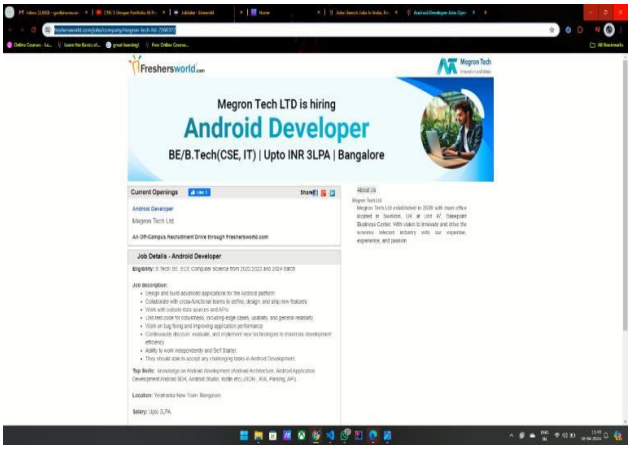



***Plot the confusion matrix(Random undersampling***

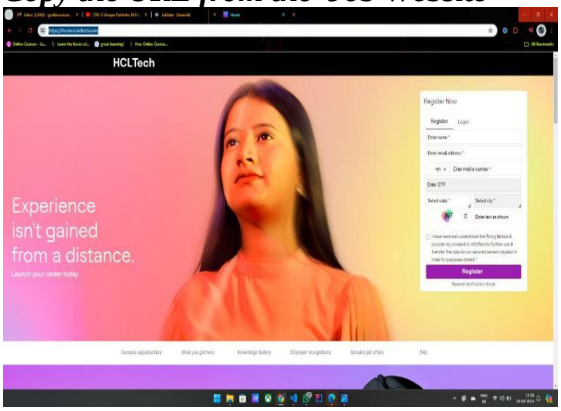



***Interface of Streamlit***

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.1, May : 2024




***Copy the URL from the Job Website***




***Copy The URL from the job website***

## CONCLUSION

In conclusion, the development of the fake job posting detection system represents a significant step towards enhancing the safety and integrity of online job searching platforms. Through the utilization of advanced machine learning techniques, data preprocessing methods, and interactive web applications, the system demonstrates promising capabilities in accurately identifying fraudulent job postings and protecting job seekers from potential scams.

The logistic regression model trained on a balanced dataset exhibits robust performance in distinguishing between fraudulent and genuine job postings, achieving high accuracy, precision, recall, and F1-score values. By leveraging text preprocessing techniques, TF-IDF vectorization, and class imbalance handling strategies, the model effectively captures important features from job postings and makes reliable predictions.

Furthermore, the deployment of the system as a Streamlit application enables real-time predictions on job postings from user-provided URLs, enhancing accessibility and usability for job seekers. Visualization techniques such as word clouds provide valuable insights into common keywords and phrases associated with fraudulent and genuine job postings, aiding in model interpretation and understanding of fraud patterns.

Accuracy: 0.9847193652659418
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.97 | 0.98 | 3396 |
| 1 | 0.97 | 1.00 | 0.98 | 3410 |
| accuracy | | | 0.98 | 6806 |
| macro avg | 0.98 | 0.98 | 0.98 | 6806 |
| weighted avg | 0.98 | 0.98 | 0.98 | 6806 |

***Accuracy and Classification***

Accuracy: 0.9164265129682997
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.93 | 0.92 | 182 |
| 1 | 0.92 | 0.90 | 0.91 | 165 |
| accuracy | | | 0.92 | 347 |
| macro avg | 0.92 | 0.92 | 0.92 | 347 |
| weighted avg | 0.92 | 0.92 | 0.92 | 347 |

***Accuracy and Classification(Random Undersampling)***

In the proposed System we got the Accuracy rate 98% where it is more than the existing system as we used the more advanced techniques to implement the project and in the real time Interface.

**Future Scope:**

While the developed fake job posting detection system demonstrates promising capabilities, there are several avenues for future research and enhancement:

1. Exploration of Advanced Models: Investigate the performance of more advanced machine learning models such as deep learning-based architectures (e.g., recurrent neural networks, transformers) for improved detection accuracy and scalability.
2. Incorporation of External Data: Integrate external data sources such as social media profiles, website reputation scores, and historical fraud databases to enrich feature representation and enhance model robustness against evolving fraud patterns.
3.Semantic Analysis: Explore semantic analysis techniques to capture deeper contextual meaning from job postings, enabling the model to identify subtle nuances and linguistic patterns indicative of fraudulent intent.
4.Real-Time Monitoring: Implement real-time monitoring and alerting mechanisms to notify users and platform administrators of suspicious job postings as they are posted, enabling proactive fraud prevention measures.
5.User Feedback Integration: Incorporate user feedback mechanisms within the application to collect insights, reports, and annotations from users regarding their experiences with job postings, facilitating continuous improvement and refinement of the model.
6.Ethical Considerations: Address ethical considerations such as privacy protection, fairness, transparency, and accountability in model development, deployment, and usage to ensure responsible and ethical AI practices. By addressing these areas of future scope, the fake job posting detection system can evolve into a robust and reliable tool for combating online job scams, safeguarding job seekers, and fostering trust and integrity in the online job marketplace.

**REFERENCES**

(1)Kang B, Kang S. Real and fake job postings classification with deep learning. In: 2018 IEEE International Conference on Big Data (Big Data). IEEE; 2018. p. 2702-2707. IEEE Xplore

(2)Aggarwal A, Gupta P, Lehal GS. Real-time fake job posting detection on social media. In: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). IEEE; 2018. p. 65-71. IEEE Xplore

(3)Selvakumar J, Nandhini B, Aravindan C, et al. Fake job postings detection using machine learning algorithm. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC). IEEE; 2018. p. 122-126. IEEE Xplore

(4)Li X, Li L, Zhu X, et al. Identifying fake online job postings by analyzing descriptions. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM; 2017. p. 2351-2354. ACM Digital Library

(5)Gupta S, Agrawal A. Fake job post detection on online job portals using machine learning. In: 2018 3rd International Conference for Convergence in Technology (I2CT). IEEE; 2018. p. 1-5. IEEE Xplore [1]