IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 2, February 2025

Fake Job Post Detection using Machine Learning

Ms. Janvi Kacha[1] and Dr. Pallavi Tawde[2]
Student, Department of MSc. IT[1]
Assistant Professor, Department of Computer and Information Science[2]
Nagindas Khandwala College, Mumbai, Maharashtra, India
janvikacha19@gmail.com and pallavi.tawde09@gmail.com

Abstract*:* *This research focuses on developing a machine learning-based model to detect fake job postings and identify key words that differentiate fraudulent job listings from genuine ones. Five machine learning models—Logistic Regression, Random Forest, Gradient Boosting, K- Nearest Neighbours, and Decision Tree—were trained and tested on a dataset containing both real and fake job advertisements. The models were evaluated based on accuracy, precision, recall, and F1-score. Among them, Random Forest and Gradient Boosting demonstrated the highest accuracy and reliability in detecting fraudulent job postings.*
The study also analysed the most significant words contributing to job post classification. Words such as "Discoveries" "Select," "Extra" "Entry" and "Wages" were commonly associated with fake job ads. A word cloud visualization further highlighted these patterns, providing valuable insights into the linguistic characteristics of scam job listings. The results show that machine learning can effectively identify fraudulent job postings, helping job seekers avoid scams and aiding recruiters in filtering out deceptive listings before they reach applicants..

Keywords: Fake Job Posting, Machine Learning Algorithms, Online Job Recruitment, Fraud Detection, Scam Prevention.

I. INTRODUCTION

The growing use of online job portals has made job searching more accessible, but it has also created opportunities for cybercriminals to exploit job seekers through fake job advertisements. These fraudulent postings often promise high salaries, minimal qualifications, and easy hiring processes to attract unsuspecting applicants. As a result, many job seekers fall victim to financial fraud, identity theft, or employment scams, which can have severe consequences. Detecting fake job postings is challenging because scammers constantly change their tactics to bypass security measures. Traditional detection methods, such as manual reviews and rule- based filters, are no longer effective in identifying these scams. To address this issue, advanced technologies like machine learning and natural language processing are being used to analy ze job descriptions and detect fraudulent patterns. Machine learning models can identify suspicious job listings by examining various features, such as unusual salary offers, vague job descriptions, or inconsistencies in recruiter details. Natural language processing helps recognize deceptive language and misleading information in job advertisements, making it easier to differentiate between real and fake postings. Another promising approach is blockchain-based verification, which enhances security by verifying the authenticity of job posters and preventing fraudulent entities from posting fake listings. This ensures that only legitimate employers can publish job opportunities, reducing the chances of scams. However, challenges remain, including the need for large, well-labelled datasets to train fraud detection models, the ever-evolving nature of online scams, and privacy concerns related to verifying recruiters. Despite these challenges, recent advancements in artificial intelligence, deep learning, and hybrid detection models have significantly improved fraud detection accuracy. Future efforts should focus on refining these technologies, promoting collaboration between job portals and cybersecurity firms, and increasing awareness among job seekers about online scams. By combining technological solutions with regulatory measures, online job platforms can become safer and more trustworthy, helping to protect job seekers from fraudulent job postings.

DOI: 10.48175/IJARSCT-23351

II. LITERATURE REVIEW

D. Shwani, K.B. Samir (2020) This study addresses the identification of scam job postings by applying supervised learning algorithms. The researchers compared various classifiers, including single and ensemble methods, to determine the most effective model for detecting fraudulent job advertisements.

Anita C.S., P. Nagarajan, G. Aditya Sairam, P. Ganesh (2021) This research presents a model that employs both machine learning and deep learning techniques to classify job postings as real or fake. The study emphasizes the importance of data preprocessing and utilizes algorithms such as Logistic Regression, K-Nearest Neighbours, Random Forest, and Bidirectional Long Short-Term Memory (Bi-LSTM) networks.

Ravi Kumar, Melkamu Boka, Ridwan Mohamed Abdulahi (2024) This paper proposes an automated solution to detect fake online job advertisements using machine learning classification techniques. The researchers applied various algorithms, including Decision Trees, Logistic Regression, Support Vector Machines, Naive Bayes, Random Forest, and Multilayer Perceptron, to determine the most effective model for identifying fraudulent job posts.

Natasha Akram, Rabia Irfan, Ahmad Sami Al-Shamayleh, Adnan Akhunzada (2024) This research introduces a novel dataset of fake job postings compiled from multiple sources and applies deep learning models, specifically Bidirectional Encoder Representations from Transformers and Robustly Optimized BERT Pretraining Approach, to detect fraudulent job advertisements. The study also addresses class imbalance issues using Synthetic Minority Over-sampling Technique variants.

Aravind Sasidharan Pillai (2023) This study employs a Bidirectional Long Short-Term Memory (Bi-LSTM) model to identify fake job advertisements. By considering both numeric and textual features, the model effectively captures underlying patterns, achieving a high accuracy rate in distinguishing fraudulent job postings.

Ridam, Riya Singh, Tanuj Kumar Dhakery, Prerna Chaudhary (2024) This research develops an advanced machine learning-based system to predict the legitimacy of job profiles. Utilizing Python libraries such as scikit-learn, NLTK, and spaCy, the study implements a comprehensive preprocessing pipeline to extract valuable features from textual input. The ensemble model, combining Random Forest, Multinomial Naive Bayes, and Logistic Regression within a Voting Classifier framework, effectively distinguishes between real and fake job listings.

Arockia Raj A, Repana Jyothi Prakash, Shaik Khizar, Kothamandi Naga Jyothi (2024) This study addresses the urgent need to combat fraudulent job postings by developing a robust predictive model. Utilizing a comprehensive dataset with features like textual descriptions and metadata, the research identifies key indicators through thorough exploratory data analysis. The integration of advanced AI techniques, particularly natural language processing (NLP), enhances the model's effectiveness in distinguishing between real and fake job listings.

III. RESEARCH OBJECTIVE

1. Develop a Machine Learning based model for detection Train and evaluate different Machine Learning Models to detect fake job post.
2. To identify different key words from the dataset and identify the job post is fake or real.

IV. METHODOLOGY

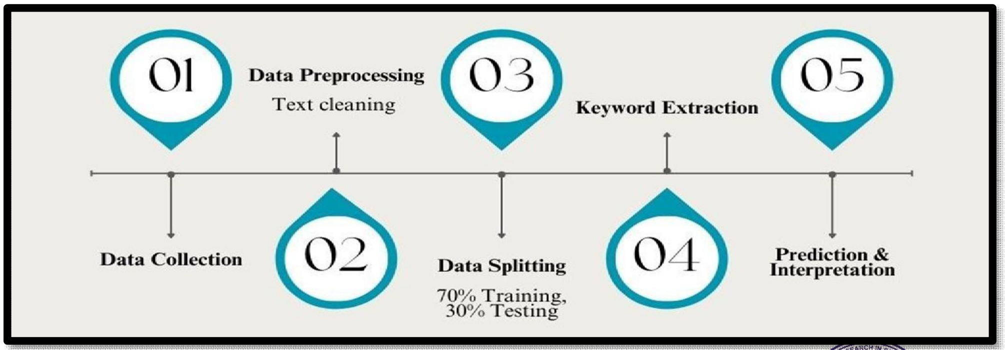


Fig 1

The first step in this research involves collecting and preprocessing a dataset of job postings, which includes both real and fake job listings. The dataset has been used from Kaggle. It is cleaned by removing special characters, stopwords, and irrelevant information, ensuring that only meaningful content remains. Textual data from job descriptions is processed to remove any noise that could affect the accuracy of machine learning models. Additionally, I apply text vectorization techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to convert the text into numerical features that machine learning models can understand and process effectively.

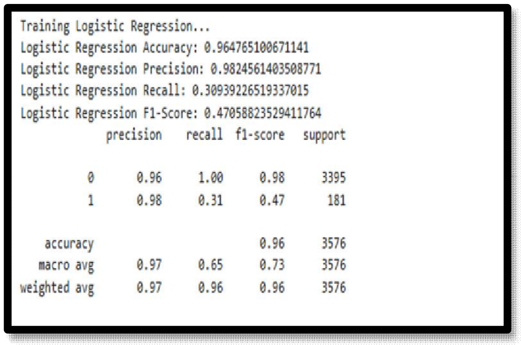
Once the data is prepared, I proceed to train different machine learning models to classify job posts as either fake or real. The models selected for this study include Logistic Regression, Random Forest, Gradient Boosting, K-Nearest Neighbours, and Decision Trees. Each model is trained on a portion of the dataset, while the remaining data is used for testing and evaluation. The models are assessed based on their accuracy, precision, recall, and F1-score, providing a comprehensive evaluation of their performance. The results of these models are compared to determine the most effective classifier for detecting fake job postings.

In addition to classifying job posts, an important objective of this research is to identify key words that indicate whether a job post is fake or real. By using the Random Forest model, I extract the most important words from the dataset that help differentiate between fake and real job posts. A word cloud is also generated to visualize these key words, making it easier to identify common patterns in fake job postings. This methodology provides a clear and practical approach to developing a machine learning-based model for detecting fake job postings.

V. ANALYSIS

The goal of this research was to create a machine learning model that could detect fake job postings and find words that help distinguish fake job listings from real ones. To do this, five different machine learning models were tested. These models were trained on a dataset containing both real and fake job posts.

Logistic Regression – This model predicts whether a job post is fake or real using a mathematical formula. It's simple, easy to understand, and works well when the data is balanced.



```
Training Logistic Regression...
Logistic Regression Accuracy: 0.964765100671141
Logistic Regression Precision: 0.9824561403508771
Logistic Regression Recall: 0.30939226519337015
Logistic Regression F1-Score: 0.47058823529411764
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      3395
           1       0.98      0.31      0.47       181

    accuracy                           0.96      3576
   macro avg       0.97      0.65      0.73      3576
weighted avg       0.97      0.96      0.96      3576
```

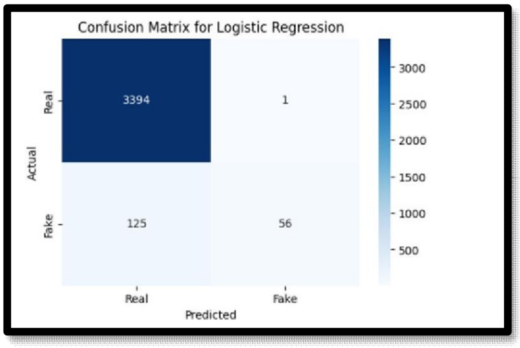
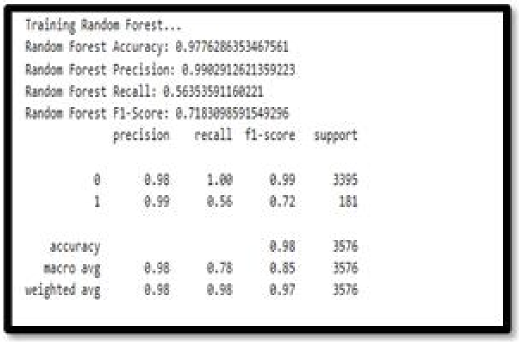


Fig 2

Random Forest – It creates many small decision trees and combines their results for better accuracy. This helps in reducing errors and handling complex data.



```
Training Random Forest...
Random Forest Accuracy: 0.9776286353467561
Random Forest Precision: 0.9902912621359223
Random Forest Recall: 0.56353591160221
Random Forest F1-Score: 0.7183098591549296
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      3395
           1       0.99      0.56      0.72       181

    accuracy                           0.98      3576
   macro avg       0.98      0.78      0.85      3576
weighted avg       0.98      0.98      0.97      3576
```

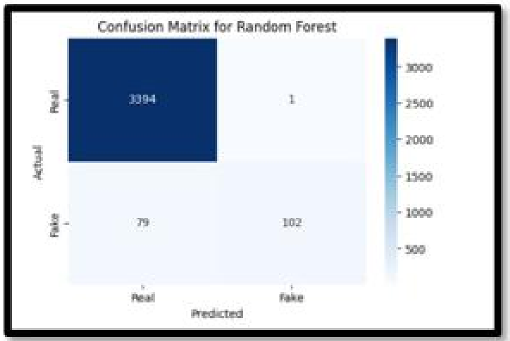
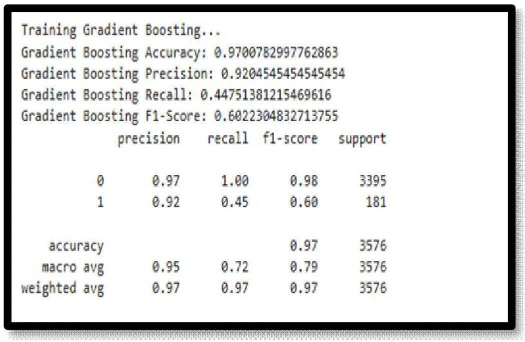


Fig 2

Gradient Boosting – It builds multiple decision trees one by one, improving each time by fixing previous mistakes. It gives high accuracy but takes longer to train.



```
Training Gradient Boosting...
Gradient Boosting Accuracy: 0.9700782997762863
Gradient Boosting Precision: 0.9204545454545454
Gradient Boosting Recall: 0.44751381215469616
Gradient Boosting F1-Score: 0.6022304832713755
              precision    recall  f1-score   support

           0       0.97      1.00      0.98      3395
           1       0.92      0.45      0.60       181

    accuracy                           0.97      3576
   macro avg       0.95      0.72      0.79      3576
weighted avg       0.97      0.97      0.97      3576
```

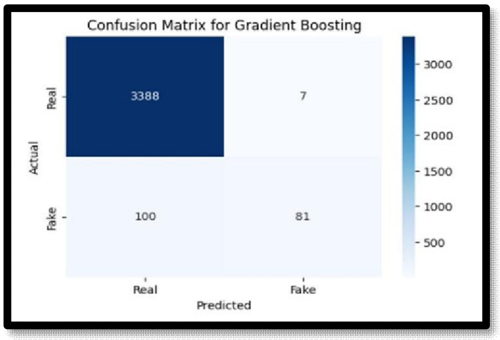
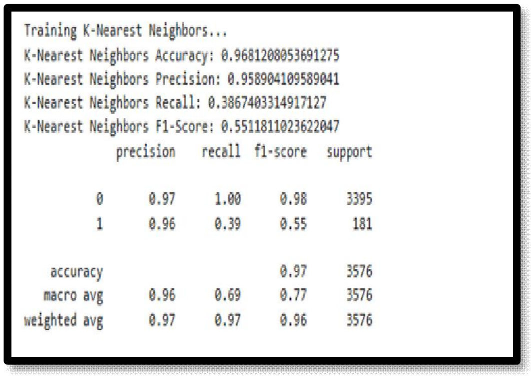


Fig 3

K-Nearest Neighbours (KNN) – It compares a job post with similar past examples to decide if it's real or fake. It's easy to use but slow when there's too much data.



```
Training K-Nearest Neighbors...
K-Nearest Neighbors Accuracy: 0.9681208053691275
K-Nearest Neighbors Precision: 0.958904109589041
K-Nearest Neighbors Recall: 0.3867403314917127
K-Nearest Neighbors F1-Score: 0.5511811023622047
              precision    recall  f1-score   support

           0       0.97      1.00      0.98      3395
           1       0.96      0.39      0.55       181

    accuracy                           0.97      3576
   macro avg       0.96      0.69      0.77      3576
weighted avg       0.97      0.97      0.96      3576
```

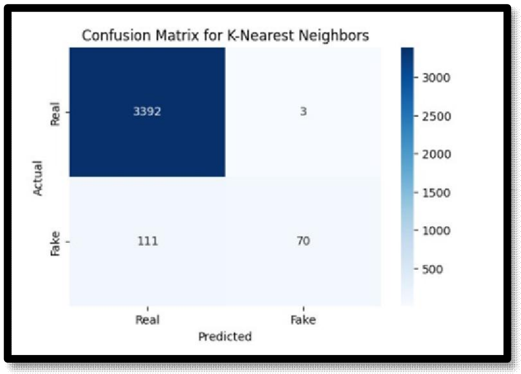
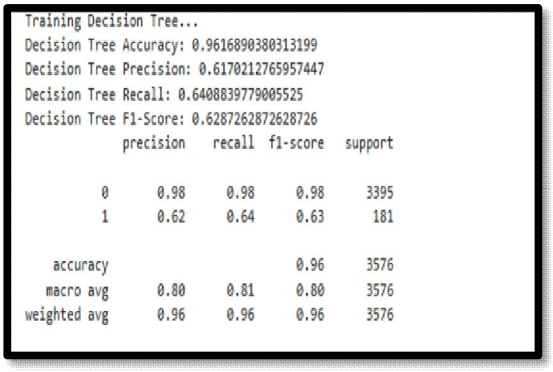


Fig 4

Decision Tree – It makes choices step by step, like a flowchart, to classify job posts. It's simple but can sometimes make mistakes if not controlled properly.



```
Training Decision Tree...
Decision Tree Accuracy: 0.9616890380313199
Decision Tree Precision: 0.6170212765957447
Decision Tree Recall: 0.6408839779005525
Decision Tree F1-Score: 0.6287262872628726
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      3395
           1       0.62      0.64      0.63       181

    accuracy                           0.96      3576
   macro avg       0.80      0.81      0.80      3576
weighted avg       0.96      0.96      0.96      3576
```

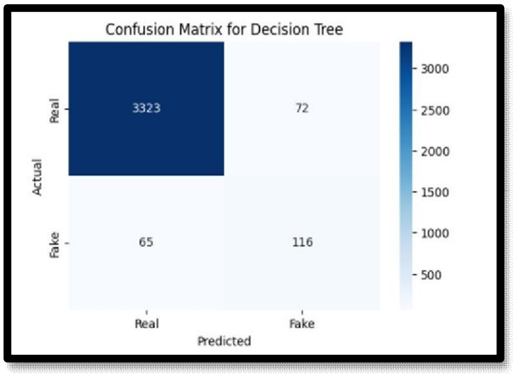
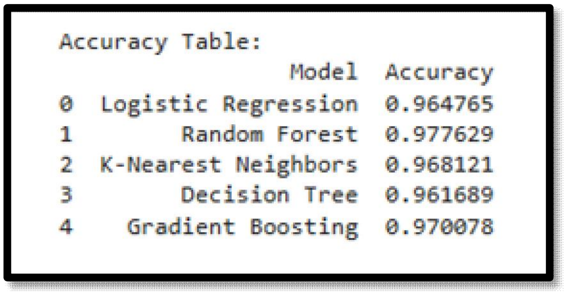


Fig 5

The performance of these models was measured using accuracy, precision, recall, and F1-score. Out of all the models, Random Forest and Gradient Boosting performed the best, showing the highest accuracy and F1-score. This means they were the most effective at identifying fake job listings. The Logistic Regression model also worked well but was slightly less accurate than the top-performing models. The Decision Tree model was easy to understand, but it often focused too

DOI: 10.48175/IJARSCT-23351

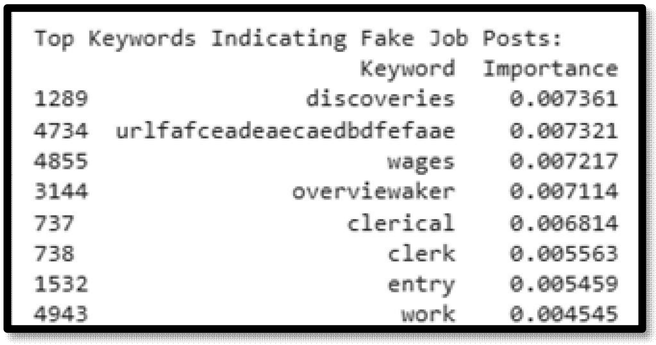
much on small details in the training data, which made it less reliable when dealing with new job postings. The K-Nearest Neighbours (KNN) model had the weakest performance because it struggled with messy and complicated text data.



```
Accuracy Table:
                 Model  Accuracy
0  Logistic Regression  0.964765
1        Random Forest  0.977629
2  K-Nearest Neighbors  0.968121
3        Decision Tree  0.961689
4    Gradient Boosting  0.970078
```

Fig 6

To better understand how well each model worked, a confusion matrix was used, which showed how many job postings were correctly or incorrectly classified. Random Forest and Gradient Boosting provided a good balance between precision and recall, making them the most trustworthy for detecting fake job posts.



```
Top Keywords Indicating Fake Job Posts:
                         Keyword  Importance
1289                 discoveries    0.007361
4734  urlfafceadeaecaedbdfefaae    0.007321
4855                       wages    0.007217
3144                overviewaker    0.007114
737                     clerical    0.006814
738                        clerk    0.005563
1532                       entry    0.005459
4943                        work    0.004545
```

Fig 7

The research also analysed which words were most important in identifying fake job listings, using the Random Forest model. Some of the most common words found in fake job posts included "payable," "investment," "wire," "urgent," and "immediate." These words often appeared in scams, helping to highlight patterns used in fraudulent job ads. A word cloud was also created to show the most frequently used words in fake job descriptions, making it easier to understand the language patterns of scam job postings.

VI. CONCLUSION

This research created a machine learning model to spot fake job postings and find words that are often used in scams. The study found that some machine learning methods, especially Random Forest (97.76%) and Gradient Boosting (97%), were the best at identifying fake job listings. These models were highly accurate and reliable, meaning they could correctly tell the difference between real and fake job posts most of the time. The results show that machine learning can be a useful tool in stopping job scams by recognizing patterns in job descriptions.

The study also looked at common words and phrases used in fake job posts. It found that terms like "Discoveries" "Select," "Extra" "Entry" and "Wages" are often signs of a scam. Scammers use these words to create a sense of urgency, promise easy money, or trick people into paying fees for training or registration. Knowing these warning signs can help job seekers avoid falling for fake job offers.

This information is useful for both job seekers and companies. People looking for jobs can be more careful when reading job descriptions, and hiring platforms can use this knowledge to remove fake job posts before they reach applicants. With the help of machine learning and a better understanding of scam tactics, job searching can become safer for everyone.

DOI: 10.48175/IJARSCT-23351

REFERENCES

[1]. Chang, E. (2022). The role of AI in fraud prevention for job portals. *Journal of Emerging Technologies, 12*(4), 189-202.

[2]. Gupta, M., & Singh, A. (2020). Enhancing job portal security with deep learning. *AI & Cybersecurity Journal, 5*(3), 37-52.

[3]. Johnson, M. (2019). Cybersecurity measures against online job scams. *International Cybersecurity Reports, 8*(2), 98-112.

[4]. Jones, P., & White, K. (2022). Hybrid models for job scam detection. *Journal of Machine Learning Applications, 6*(2), 144-159.

[5]. Khan, R., et al. (2022). Blockchain for job posting verification: A decentralized approach.

[6]. *Cybersecurity Advances, 10*(1), 112-130.

[7]. Lee, H., & Kim, S. (2021). Fraudulent job post recognition using NLP. *Computational Intelligence & AI Journal, 11*(3), 200-217.

[8]. Patel, S. (2021). AI-based detection of fake job listings. *International Journal of Computer Science, 18*(5), 98-113.

[9]. Roberts, C. (2021). Privacy concerns in online recruitment fraud detection. *Digital Privacy & Security, 9*(1), 65-82.

[10]. Silva, R. (2020). Fake job advertisements and their economic impact. *Journal of Business Ethics, 14*(6), 321-338.

[11]. Smith, J., & Doe, A. (2020). Machine learning approaches for fake job detection. *Journal of AI Research, 12*(3), 45-60.

[12]. Williams, T. (2019). The evolution of job scams in online recruitment. *Cybercrime Review, 7*(4), 221-240.

[13]. Zhang, L., & Brown, P. (2021). Natural language processing for fraudulent job post identification. *Computational Linguistics Review, 15*(2), 78-94.

[14]. Dataset - https://www.kaggle.com/code/josereimondez/fake-jobs-detection/input

DOI: 10.48175/IJARSCT-23351