IJMRSET

e-ISSN:2582-7219

# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

## IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 7, July 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707153 |

# Assessing Different Data Mining Techniques for Predicting Fake Job Listings: "A Comparative Analysis"

**Prof. Sravanthi, Shashank H S**

Assistant Professor, Department of MCA, AMC Engineering College, Bengaluru, India

Student, Department of MCA, AMC Engineering College, Bengaluru, India

**ABSTRACT:** Fake job postings are a growing concern market, deceiving job seekers and causing financial and personal harm.
To find the best strategy, this study evaluates many data mining strategies for identifying phony job postings. Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Naive Bayes, k-nearest Neighbors, and Neural Networks are some of the techniques evaluated. We assess their effectiveness using criteria such as recall, accuracy, and precision.
F1 score, and ROC-AUC on a dataset of actual and false job postings. According to our findings, ensemble approaches—Random Forests in particular—Please remember the following text: TEXT: outperforms other methods in terms of resilience. And accuracy, making them a dependable option for identifying phony job listings.

**KEYWORDS:** "Fake Job Postings, Data Mining, Certainly! Here is the revised text: "Logistic Regression, Decision Trees, Random Forests, Support Vector Machines" Bayes, k-Nearest Neighbors, Neural Networks

## I. INTRODUCTION

The increase in online job portals has resulted in more fake job postings. These fraudulent listings can cause serious financial and personal harm to job seekers. It is crucial to detect fake job posts to protect users and maintain the credibility of online job platforms. Data mining methods offer effective ways to job postings. This article compares various data mining techniques used to tackle the issue of detecting fake job posts,

highlighting their strengths, weaknesses, and overall effectiveness."

### 1.1 Problem Statement

Fake job postings can be misleading, wasting job seekers' time and potentially leading to identity theft or financial fraud. Despite advancements Such postings are sophisticated and ever-evolving, making machine learning detection of them a difficult challenge. and Neural Networks are among the methods assessed. We evaluate their efficacy based on parameters including recall, accuracy, and precision
to assess various data mining methods. For their efficacy in identifying fake job posts, and providing best practices for enhancing detection systems.

### 1.2 Objectives

To evaluate the performance of various data mining techniques in identifying fraudulent job postings. To determine the most effective methods based on multiple performance measures. To offer suggestions for enhancing systems for detecting fake job postings.

## II. LITERATURE REVIEW

### 2.1 Fake Job Post Detection

The detection of fake job postings has received a lot of attention in recent years. Previous research has employed a range of methods, from rule-based approaches to advanced machine-learning techniques. For example, Bhatia et al. (2016) used a rule-based approach to identify fake job advertisements based on specific trends and phrases. Narayan and Joshi (2018) explored the use of Naive Bayes and other neural network classifiers, as well as decision trees, and achieved moderate success in distinguishing between fake and real job posts.

### 2.2 Data Mining Techniques

- Here is the revised text:

A binary classifier is used in the statistical technique of logistic regression to represent the likelihood of a default group. Decision Trees use a tree-structured paradigm to divide data into subgroups based on attribute values. "Random Forests is a method that creates several decision trees in order to make predictions." decision trees and combines them to improve forecasting accuracy. Support Vector Machines (SVM) is a supervised learning model that identifies the optimal plane to separate classes. Naive Bayes is a statistical method that applies strong independence assumptions across features of the Bayes principle. K-Nearest Neighbors (k-NN) is a non-parametric technique that classifies examples based on the majority class among their closest neighbours.

- Neural Networks Models of artificial intelligence are created to resemble the human brain.

by utilizing interconnected layers of neurons to recognize complex patterns. Comparative studies.

Virtual learning like Rama and Ho (2020) highlights that ensemble methods, particularly Random Forests, often outperform single models in various classification tasks. Studies such. (2019) have demonstrated the efficacy of SVM and Neural Networks for identifying irregularities throughout huge data sets.

## III. METHODOLOGY

### 3.1 Dataset

The dataset used to investigation contains job postings scraped from various online platforms, labeled as either real or fake. It includes attributes such as job title, company, location, salary, and description.

3.1.1 **Total Posts**: 20,000
3.1.2 **Real Posts**: 15,000
3.1.3 **Fake Posts**: 5,000

### 3.2 Preprocessing

Incorrect value handling, categorical attribute storage, and statistical feature normalization were all as a component of the data processing. Text data, such as job descriptions, were converted into vectors using TF-IDF (Term Frequency-Inverse Document Frequency).

### 3.3 Techniques

3.3.1 **Logistic Regression**: Implemented using scikit-learn with L2 regularization.
3.3.2 **Decision Trees**: Gini index used for splitting, with a maximum depth of 10.
3.3.3 **Random Forests**: 100 trees with bootstrapping.
3.3.4 **SVM**: RBF kernel with a penalty parameter of C=1.0.
3.3.5 **Naive Bayes**: Multinomial distribution applied to text features.
3.3.6 **k-NN**: k=5 with Euclidean distance.
3.3.7 **Neural Networks**: Three buried sheets with ReLU activation, trained with Adam optimizer.

### 3.4 Evaluation Metrics

3.4.1 **Accuracy**: (TP + TN) / (TP + TN + FP + FN)
3.4.2 **Precision**: TP / (TP + FP)

3.4.3 **Recall**: TP / (TP + FN)
3.4.4 **F1 Score**: 2 * (Precision * Recall) / (Precision + Recall)
**ROC-AUC**: Region falling inside the Receivers Operation Parameter curve.

## IV. RESULTS AND DISCUSSION

### 4.1 Performance Metrics

Each data-exploitation technique's efficacy was assessed using the metrics mentioned above. Table 1 summarizes the results.

**Table 1: Performance Metrics**

| **Technique** | **Accuracy** | **Precision** | **Reca ll** | **F1 Scor e** | **RO C-AU C** |
|---|---|---|---|---|---|
| Logistic Regression | 85.2% | 84.7% | 83.5 % | 84.1 % | 0.89 |
| Decision n Trees | 87.3% | 86.9% | 85.8 % | 86.3 % | 0.90 |
| Random Forests | **91.5%** | **91.0%** | **90.2 %** | **90.6 %** | **0.95** |
| SVM | 88.7% | 88.0% | 87.5 % | 87.8 % | 0.93 |
| Naive Bayes | 82.1% | 80.5% | 79.8 % | 80.1 % | 0.86 |
| k-NN | 84.5% | 83.0% | 82.4 % | 82.7 % | 0.88 |
| Neural Networ ks | 89.2% | 88.5% | 87.9 % | 88.2 % | 0.92 |

## V. DISCUSSION

. Recursive forest construction produced the greatest results in terms of accuracy, 8 recall, F1 rating, precision, and ROC-AUC. The ensemble nature of Random Forests minimizes overfitting and helps capture complex patterns.
Additionally, artificial neural networks and Performance of support vector machine models was similarly good. demonstrating their ability to handle non-linear interactions in the data.

- **Decision Trees** provided a clever evaluation of interpretability and performance but were outperformed by their ensemble counterpart, Random Forests.

**LogisticRegression** , **k-NN** showed reasonable performance but were less actual likened to the ensembleand more complex models.

- **Naive Bayes** had lowest performance, likely unpaid to the strong independence assumptions that do not hold well in the perspective of fake job postings.

## VI. CONCLUSION

This comparative study evaluated many data removal methods for detecting fake job posts, finding that ensemble methods, particularly Random Forests, provide superior performance. Encouragement Neural network techniques produced excellent results as well; basic Additionally successful were support vector machine models.
, and naive bayes, performed worse. The results indicate that utilizing group techniques and powerful machine learning models can significantly enhance the detection of fraudulent job postings.

## VII. FUTURE ENHANCEMENTS

1. **Feature Engineering**: Explore additional features such as user behavior data and job posting metadata to improve detection accuracy.
2. **Hybrid Models**: Develop hybrid models combining multiple techniques to leverage their individual strengths.
3. **Real-Time Detection**: Optimize models for real-time detection to provide immediate alerts on fraudulent job postings.
4. **Explainability**: Incorporate model interpretability techniques to provide insights into decision-making processes, enhancing trust and usability.
5. **Adaptation**: Implement adaptive learning mechanisms to continually update models based on new data and emerging patterns.

## REFERENCES

1. Bhatia, P., Jain, S., & Gupta, M. (2016). *A study of fake job advertisement detection techniques*. IEEE International Conference on Big Data, 755-760. doi:10.1109/BigData.2016.7556788
2. Rama, K. & Ho, J. (2020). *Ensemble methods for classification and their applications*. arXiv preprint arXiv:2004.09633.