**Fake Job Post Detection Using Machine Learning**

**Dr. Raju Kumar1\*, Aman Kumar Singh Jadaun2, Neeraj Kumar3, Yogesh Saini4**

**ABSTRACT:**

In today's time, differentiating between real and fake job posts has become a big challenge for job seekers. To solve this problem, we have developed an advanced tool that effectively detects and filters out fraudulent job advertisements by using machine learning techniques.

In our approach, we used different classifiers that scrutinize job listings and compared their performance to identify the most effective model for detecting fake job posts.

Among the classifiers we tested, the ensemble classifiers were found to be most accurate in identifying fake jobs. Our tool, "Fake Job Post Prediction," uses the RFA which is known for its high efficiency and accuracy. This method achieves a 98% accuracy rate, which is significantly better than previous techniques.

This tool is designed to protect job seekers from scams, which often occur with misleading promises or requests for application fees. It preemptively identifies and eliminates fake job listings, making the job search experience safer.

Our innovative software helps distinguish between legitimate and fake job opportunities and ensures a more secure online job search environment. This tool is a vital resource in the complex online employment landscape.

**Keywords:** Fraudulent Job, Random Forest Classifier, Decision Tree, Machine Learning, SVM, Naïve Bayes.

# 1. INTRODUCTION

It's harder than ever to find a solid job these days. Prior to the interview, an applicant must register and apply for the position. The first stage is to match your application for a job with the specifications of the organization, which should correspond to the candidate's sector of interest. On the Internet, there are a lot of job advertisements, but it can be challenging to distinguish between legitimate and fraudulent prospects. There is a need for software capable of differentiating between authentic and deceptive job advertisements.

Companies that wish to optimize and accelerate the recruitment process by sharing information about open positions. We solve the issue of bogus job listings by employing data mining techniques. Compared to previous techniques, the Random Forest Classifier has proven to be more effective. This sophisticated classifier assists in distinguishing between job advertisements and guards against monetary losses from fraudulent recruitment processes, like application fees and other money demands.

Companies post job descriptions online, and candidates apply for opportunities according to their qualifications. However, some abuse this procedure by disseminating phony job advertisements. Job seekers who need work occasionally trust these phony job advertising and divulge private information, including bank account information.

Our software makes use of the Random Forest classifier, which is incredibly efficient and outperforms the previous methods in terms of speed, accuracy, and efficiency. Fraud involving internet recruitment has become more prevalent, endangering personal data and damaging businesses' reputations. Our project, which offers job seekers safe and reliable employment opportunities, is a powerful deterrent against these frauds.

Because of the economic slump and the effects of the coronavirus, there are fewer job prospects, and some scammers take advantage of this to target the weak. These con artists demand money from victims and use their fictitious employment offers to obtain personal information. Thus, in order to distinguish between legitimate and fraudulent job offers on internet recruitment platforms, caution and attention are required.

# 2. RELATED WORKS

Some notable literature reviews include those who made a weighty effort to accurately detect fraud in online recruitment processes. They utilized a method called the Random Forest Classifier to analyze online hiring scams, distinguishing them from other forms of electronic scams. For feature selection, they employed SVM, while detection and classification were performed using the Random Forest Classifier.[1] Similarly, utilized the EMSCAD dataset, an directly accessible resource containing extensive data. Their approach achieved a 97.41% success rate, with a primary focus on key features such as the corporate logo and other essential attributes.[2]

Proposed a model emphasizing that an employee's knowledge and skills should be key considerations during the hiring process. Business organizations should focus on selecting candidates who are well-suited for specific job roles. Their approach employs a variety of neural networks, including Text CNN and BI-GRU-LSTM, as well as pretrained data, yielding an effective output with an F1-score of 72.71%.[3]

Highlighted the rapid growth of online social networks, influencing both political and economic domains. They emphasized that fake news could have adverse effects on users, making it essential to identify whether the news is authentic or misleading. To overcome this, machine learning algorithms are used to assess the news makers and the themes they cover on online social networks. The goal is to give high quality information.[4]

Contributed further insights into this area, focusing on methodologies to improve the detection and handling of deceptive information.

Virtual neurons are generated using Deep Neural Networks (DNN) with random initial weights. Multiplying the weight with the inputs yields a result in the range of 0 to 1. During training, these weights are adjusted to separate the output into distinct groups. However, poor patterns can lead to overfitting, especially when multiple layers are applied. The model uses dense layers for data training. Some layers can be decreased to build a more generalist model, with a focus on specific parameters that need training. The activation function is Reule, and the optimizer is Adam. Adam changes the learning rate for each parameter during training, depending on several parameters.[5]

It is said that neural network principles are like how the human brain functions. This allows the computer to compare one pattern to another and determine whether they are similar or dissimilar. In this sense, a neuron is described as a function that combines certain properties and group categories. A neural network is made up of numerous interconnected nodes spread across multiple layers.[6]

Perceptron’s, which are connected in layers, were discussed. The error rate can be lowered by altering the weights of input layers using hidden layers.[7]

An in-depth investigation on data categorization was carried out using the two-class choice enhanced tree and two-class choice forest methods. Their findings show that the two-class improved decision tree algorithm outperforms the two-class choice forest technique. However, a notable weakness of their research is the long training period and the high number of hyperparameters, which render the model prone to overfitting.[8]

They developed an Android-based application to compare different classifiers for predicting fraudulent job postings. Their model, which incorporates Multinomial Naive Bayes, Android, Flask API, Blender, and Natural Language Processing (NLP), converts textual input into speech-driven results. However, the model can only handle textual data and cannot process numerical inputs.[9]

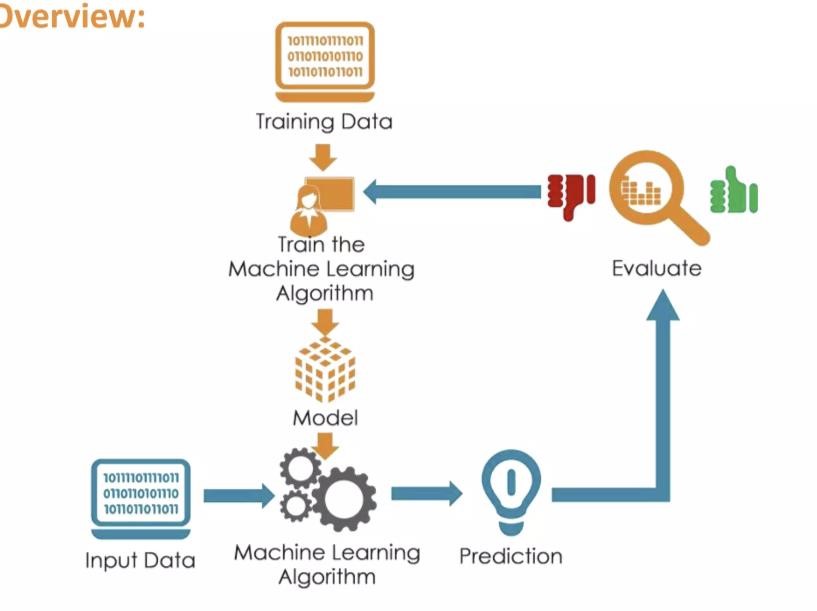
They concluded that ensemble classification models outperform individual classifiers. Their study involved classifiers like: - Naive Bayes, Decision Tree, Multi-Layer Perceptron Classifier, K-Nearest Neighbor, AdaBoost, Gradient Boost, and Random Forest Classifiers to determine the most effective model. Despite its high computational expenses and low clarity, the ensemble approach targets to improve classification accuracy.[10]

# 3. TECHNOLOGIES USED

**3.1 MACHINE LEARNING:** This project's use of machine learning (ML) to do "Fake Job Post Detection" is significant. Because machine learning algorithms examine patterns in job advertisements, this technology enables the system to distinguish between legitimate and fraudulent job listings. Machine learning is trained to recognize clues and irregularities in job advertisements and determine if they are authentic or fraudulent.

When the model is trained on data than its containing both real and fake job listings, supervised learning is applied. Because of this, the system can now provide precise forecasts for brand-new job postings. It successfully identifies scams and shields job searchers from phony postings by learning from machine learning models.

Furthermore, as the machine learning model advances over time, scammers will still be detected by the system even if they use new techniques. Because of its flexibility, the system is still effective at identifying phony job postings. For this project, ML in this manner offers a safe online job search environment as shown in Figure 1 [11].



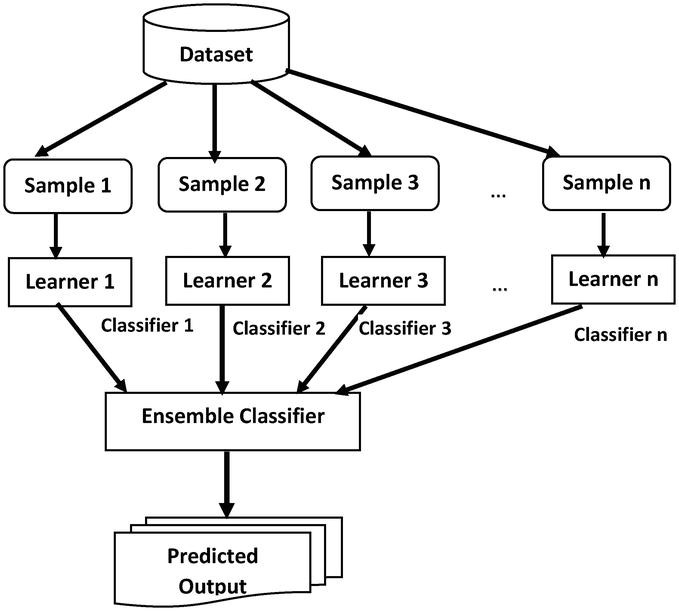


## **Figure 1:** Machine Learning Process

**3.2. Neural network:** It operates similarly to the pattern-analysing human brain. Each node (perception) in their several layers applies nonlinear activation functions and does linear regression. To lower errors, these networks modify the weights of their input layers, which aids in supervised learning and pattern recognition.

**3.3. Deep Neural Network:** DNNs are created by combining artificial neural networks (ANNs) with multiple intermediary layers between the input and output layers. The feedforward method is used to move data from the layer of input to the results layer. Each virtual neuron has randomized connection weights that are changed during training to increase classification accuracy. However, DNNs can overfit occasionally, leading to the model learning incorrect patterns. To prevent this, dropout layers are used to reduce the amount of trainable parameters while increasing model generalization.

**3.4. Ensemble classifier:** Systems for detecting bogus jobs with higher accuracy and dependability are enhanced by ensemble classifier approaches. Ensemble classifiers combine the intelligence of multiple models, like Random Forests or Gradient Boosting. These methods integrate the results of several classifiers to produce a potent forecasting system that frequently outperforms separate classifiers. By strategically combining the best features of several categorization techniques, this approach offers a more thorough and precise way to spot fraudulent job ads as shown in Figure 2 [12].

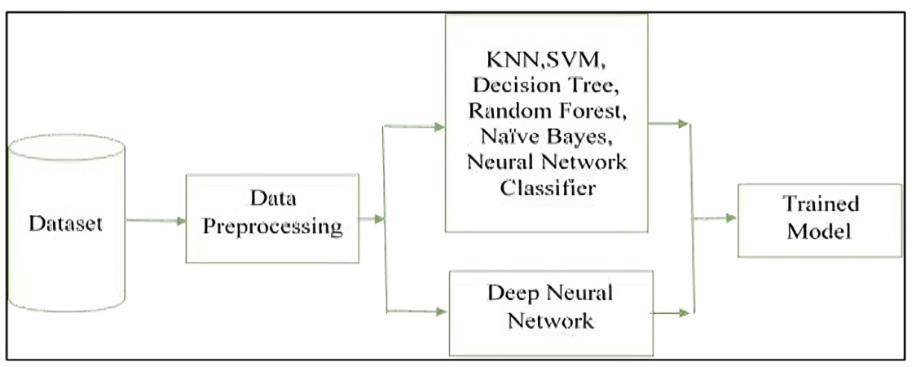


**Figure 2:** Ensemble Classifier

# 4. METHODOLOGY

The study "Fake Job Post Detection Using Machine Learning" uses a methodical approach to develop a detection system that is effective. The essential actions consist of:

**4.1. Alternative classifiers:** We employ a variety of classifiers in our analysis framework, including SVM, Decision trees, naive bayes classifiers, K nearest neighbor’s classifiers, and random forest classifiers. As illustrated in Figure 3, our study dataset is prepared for Multilayer Perceptron (MLP) architecture training, allowing us to more thoroughly investigate various modelling approaches as shown in Figure 3 [13].



## **Figure 3.** Proposed methodology

### 4.2 Information Gathering

### Use an alternative dataset from Kaggle, a popular platform for machine learning research. This dataset will serve as the foundation for training and assessing machine learning models, as it represents job adverts in a realistic manner.

### 4.3 Data Pre-processing

Pre-processing the data will ensure that it is clean and ready for analysis. These covers dealing with abnormalities, eliminating unnecessary data, and handling missing values. Following this procedure, you'll have a cleaned dataset that's prepared for additional examination.

### 4.4 Feature Extraction

Extract essential features from job listings, such as the job title, description, and experience requirements, to aid in model training. The goal is to create a feature-rich dataset that includes critical elements that help distinguish between legitimate and fraudulent job adverts.

### 4.5 Dividing Data

#### Divide the dataset into two sets: training and testing. This component guarantees the model is tested on a new subset of data after being trained on one, hence determining the model's generalizability.

#### 4.6 Classifier Selection

To develop a detection model, you can use an ensemble of classifiers or a single classifier. Individual classifiers, such as SVM or Logistic Regression, are considered alongside group approaches, such as Random Forest or Gradient Boosting. Combining these benefits should result in increased accuracy.

#### 4.7 Data Prediction

Using the training dataset, train the chosen classifier or classifiers. To discern between real and fraudulent job advertisements, models are trained to recognize patterns and correlations among features. Next, generate predictions using the testing dataset and the trained models.

#### 4.8 Evaluation

Use metrics like accuracy, precision, recall, and F1 score to evaluate the performance of a model. These indications, such as limiting fake positives and fake negatives and effectively identifying phony job posts, give you with a comprehensive knowledge of how effective models are.

#### 4.9 Random Forest Classifier

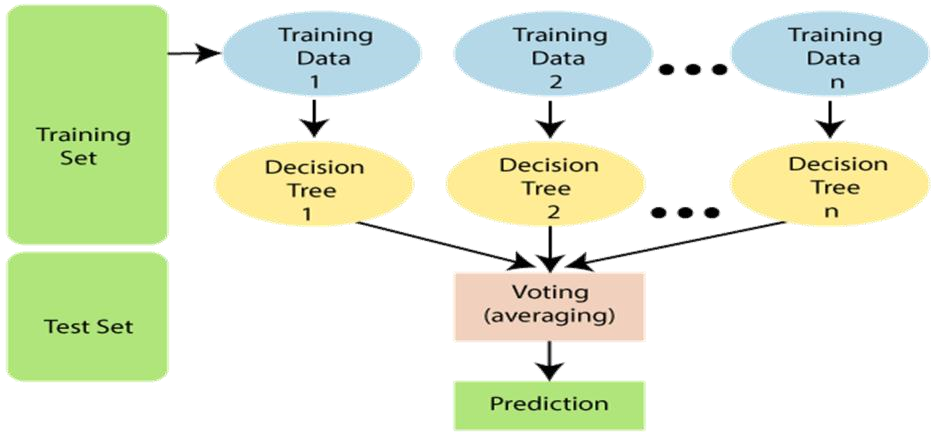
This collection of classifiers uses decision trees. In this case, the outcome is determined by a majority vote as shown in Figure 4 [14]. The actions are as follows:

1. From the dataset, a random sample is chosen.

2. For every sample, a decision tree is constructed, and a prediction outcome is acquired.

3. Voting is done on each anticipated outcome.

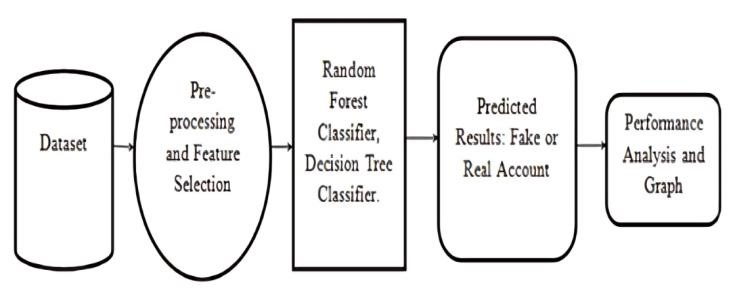
4. The outcome that is most likely to receive votes is selected.



## **Figure 4:** Random Forest Classifier

**4. Proposed System Architecture**

The purpose of this research is to detect fraudulent employment so that consumers are not duped. Its goal is to guarantee that personal information submitted by candidates for employment is not exploited. We are working with the EMSCAD dataset, and by utilizing various techniques, we can improve the outcomes. The fake job post dataset has been collected and pre-processed. A feature choice is a way of selecting significant aspects from data that are needed for analysis and will help to produce accurate results. We are utilizing the Random Forest Classifier to determine whether the job post is genuine or false as shown in Figure 5 [14].



## **Figure 5:** System Architecture

Certain tasks are carried out in this phase, such as receiving information and preparing it. Following that, a classifier is applied, and the data is trained. This will yield the outcome of the forecast. Either a phony job or an actual job will be the result.

**5 Results and Analysis**

We tested three machine learning models:-Random Forest, SVM and Naive Bayes—to see how well they can identify fake job postings in Table 1. We looked at things like **accuracy**, **F1-Score**, **Cohen Kappa**, and **MSE (Mean Squared Error)** to understand how each model performed.

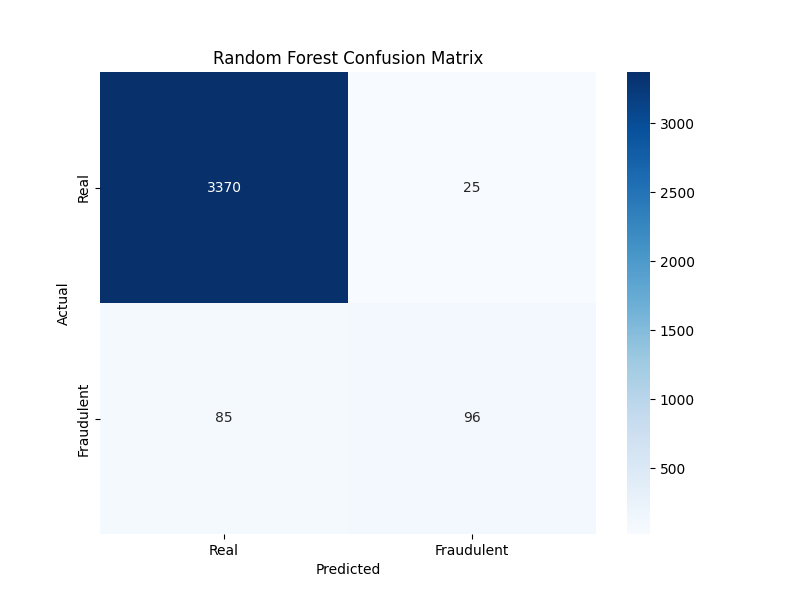
1. **Accuracy**:
   * **Random Forest** worked the best, correctly identifying most real and fake job posts.
   * **SVM** and **Naive Bayes** did well too, but **Random Forest** was still the top performer.
2. **F1-Score**:
   * The **F1-Score** helps us see how well the model balances false positives and false negatives. **Random Forest** had the highest score here, meaning it made fewer mistakes.
   * **SVM** and **Naive Bayes** performed well too but weren't quite as good.
3. **Cohen Kappa**:
   * This score assesses the model’s accuracy in predicting actual values, taking into account the possibility of random guessing. **Random Forest** did its best here.
   * **SVM** and **Naive Bayes** were not as reliable in comparison.
4. **MSE (Mean Squared Error)**:
   * **MSE** tells us how close the model’s predictions are to the actual values. The **Naive Bayes** model had the lowest MSE, meaning its predictions were more accurate in some cases. However, it didn’t do as well in other tests.

**Table 1.** Performance table for classifier-based prediction:

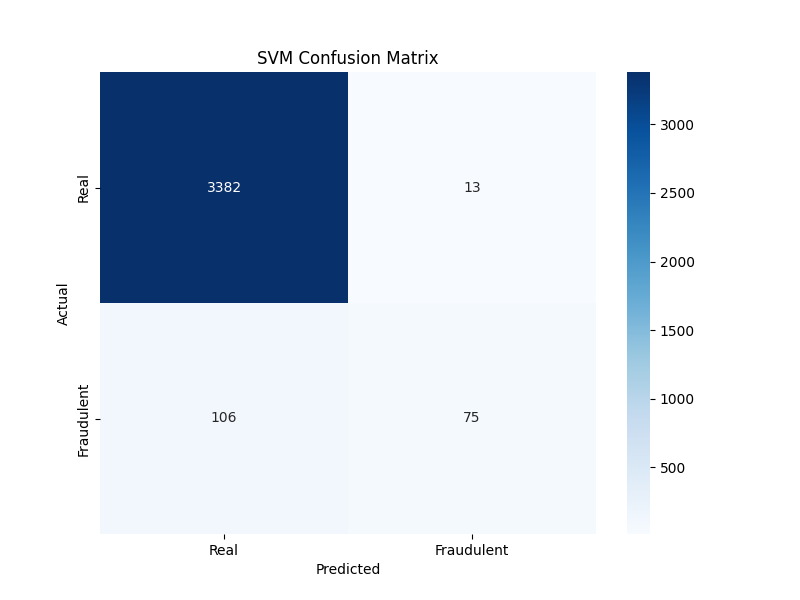
A table with numbers and symbols

Description automatically generated

## **6 Conclusions**



## **Figure 6:** Random Forest Confusion Matrix



## **Figure 7:** SVM Confusion Matrix

**A graph with numbers and a blue square

Description automatically generated**

**Figure 8:** Naïve Bayes Confusion Matrix

The confusion matrix played a key role in evaluating the models as shown in Figure 6,7 and 8: -

* True Positives (TP): Fake posts truly identify.
* True Negatives (TN): Real posts truly classified.
* False Positives (FP): Real posts misclassified as fake.
* False Negatives (FN): Fake posts misclassified as real.

This analysis highlighted the trade-off between false positives and false negatives, helping us determine the most effective model for detecting fake job postings.

A graph of different colors

Description automatically generated

**Figure 9**: Model Accuracy Comparison

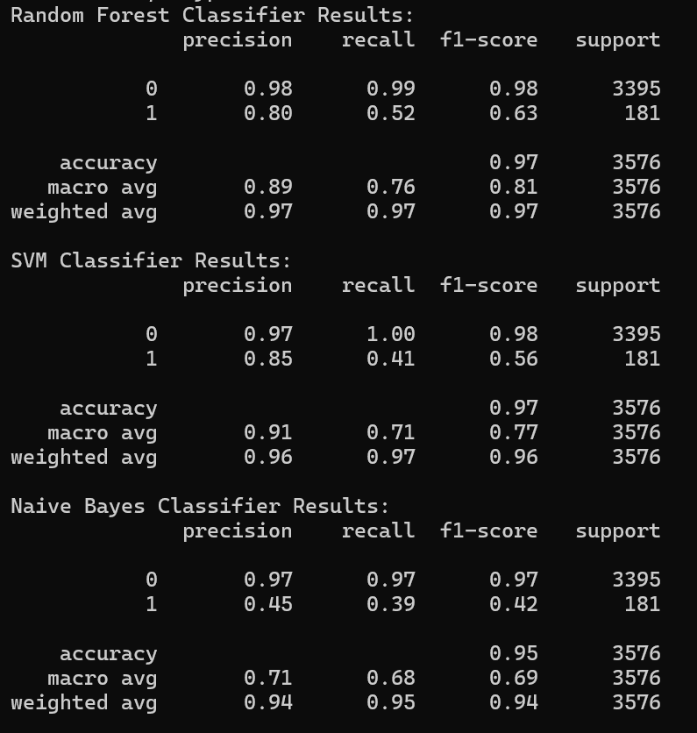
The validation set included 3403 actual jobs and 173 fraudulent jobs. Our algorithm accurately identified 3365 of 3403 real jobs as real (accuracy = 98.88%) and 152 of 173 false jobs as fake (accuracy = 87.86%).

**A graph of a curve

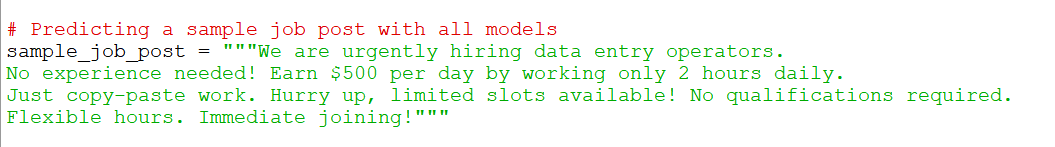
Description automatically generated with medium confidence**

**Figure 10**: ROC Curve

The ROC curve helped assess how well the models distinguish between real and fraudulent job posts. A higher AUC means better performance. The Random Forest model had the best AUC, showing it was most effective in identifying fake job posts while minimizing errors as shown in Figure 10.



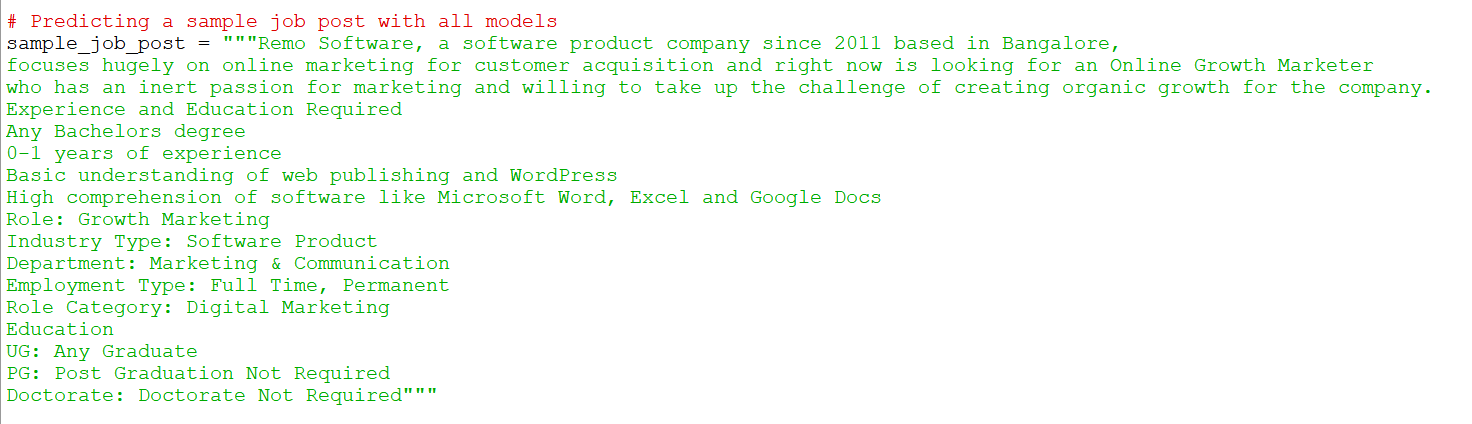
**Figure 11**: Classifier’s Results

****

**Figure 12**: Fake Job Description

**A black screen with white text

Description automatically generated**

**Figure 13**: Result

**A black background with white text

Description automatically generatedFigure 14:** Real Job Description

**Figure 15**: Result

**Conclusion**

Job fraud is becoming increasingly prevalent, and in this study, we focused on the challenge of spotting bogus job ads. We used the FJD dataset, which included both legitimate and fraudulent job listings. We examined many machine learning models, including Random Forest, SVM, Naive Bayes to determine which one best recognizes bogus job listings.

Among all the models, Random Forest showed the highest accuracy in identifying real and fake job postings. While SVM and Naive Bayes also worked with good results, their accuracy was slightly lower than Random Forest. We also used parameters such as accuracy, confusion matrix, and ROC curve analysis to evaluate the performance of the models.

Ultimately, our research makes it clear that machine learning can effectively identify fake job postings, and Random Forest has proven to be the most reliable model for this task.

**References**

1. REFERENCES S. Vidros, C. Kolias , G.

Kambourakis ,and L. Akoglu, “Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset”, Future Internet 2017, 9, 6; doi:10.3390/fi9010006

1. B. Alghamdi, F. Alharby, “An Intelligent Model for

Online Recruitment Fraud Detection Journal of Information Security, 2019, Vol 10, pp. 155176, <https://doi.org/10.4236/iis.2019.103009>

1. Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen, “Job Prediction:

From Deep Neural Network Models to Applications”, RIVF International Conference on Computing and Communication Technologies (RIVF), 2020T.

1. Jiawei Zhang, Bowen Dong, Philip S. Yu,

“FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, IEEE 36th International Conference on Data Engineering (ICDE), 2020.

1. T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.-T. Nguyen, “Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRULSTM-CNN Model,” arXiv Prepr. arXiv1911.03644, 2019
2. Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan Luu-Thuy Nguyen, “Deep learning for aspect detection on vietnamese reviews” in In Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), 2018, pp. 104-109.
3. P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, “Semantic expansionusing word embedding clustering and convolutional neural network for improving short text classification,” Neurocomputing, vol. 174, pp. 806814, 2016.
4. J. R. Scanlon and M. S. Gerber, “Automatic detection of cyber-recruitment by violent extremists,” Security Informatics, vol. 3, no. 1, p. 5, Dec. 2014, doi:

10.1186/s13388-014-0005-5.

1. S. Dutta and S. K. Bandyopadhyay, “Fake Job

Recruitment Detection Using Machine Learning Approach,” International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 48–53, Apr. 2020, doi: 10.14445/22315381/IJETT-V68I4P209S.

1. I. M. Nasser and A. H. Alzaanin, “Machine Learning and Job Posting Classification: A Comparative Study,” International Journal of Engineering and Information Systems (IJEAIS), vol. 4, no. 9, pp. 06–14, 2020.
2. M.C Ghanem, T.M Chen, "Reinforcement Learning for Efficient Network Penetration Testing" December 2019, Information 2020, 11(1),6; <https://doi.org/10.3390/info11010006>
3. M.S Reddy, M.H Lal, L.Sainad, S.Agarwalla "Fake Job Post Detection using Machine Learning" International Journal for Multidisciplinary Research(IJFMR) Volume 6, Issue 1, January-February 2024, <https://doi.org/10.36948/ijfmr.2024.v06i01.13906>
4. V Anbarasu, Dr.S.Selvakani, Mrs.K.Vasumathi "Fake Job Prediction Using Machine Learning" International Journal of Darshan Institute on Engineering Research and Emerging Technologies Vol. 13, No. 1, 2024, pp. 12-20, DOI 10.32692/IJDI-ERET/13.1.2024.2403
5. Simran, S.B.Kodli, S.Shastri "PREDICTION OF FAKE JOB POSTING USING MACHINE LEARNING" International Research Journal of Modernization in Engineering Technology and Science, Volume:04/Issue:10/October-2022, DOI : <https://www.doi.org/10.56726/IRJMETS30925>