ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

Fake Job Post Prediction Using Machine Learning Techniques

P.Narasimhaiah[1], JakkalaMamatha[2], Annapureddy Jahnavi[3],Chimmani Triveni[4], GongadiJahnavi[5], KatamLahari[6]
Chaitanya Bharathi Institute of Technology, Proddatur, A.P, India.

narasimhareddypolu@gmail.com[1],jjakkalamamatha@gmail.com[2],jahnaviannapu@gmail.com[3],chimmanitriveni7@gmail.com[4],gongadijahnavi@gmail.com[5],katamlahari@gmail.com[6]

ABSTRACT

In recent years, due to advancement in modern technology and social communication, advertising fake job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naïve bayes classifier, random forest classifier, multilayer perceptron and deep neural network to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD)containing 18000 sample.Deep neural network as a classifier, performs great for the classification task.The trained classifier shows approximately 98% classification accuracy(DNN) to predict a fraudulent job post.

Keywords: Fake Job Post Detection, Machine Learning, Online Recruitment Fraud, Random Forest Classifier, Deep Neural Network (DNN), Employment Scam Aegean Dataset (EMSCAD).

I. INTRODUCTION

The rise of digital platforms and modern communication technologies has revolutionized the way job opportunities are advertised and accessed. Job seekers now rely heavily on online platforms and social media to find employment, providing employers with a convenient way to post job vacancies. However, this shift has also given rise to a growing number of fraudulent job postings, commonly referred to as "employment scams." These scams often aim to exploit job seekers by offering fake job opportunities, stealing personal information, or demanding payments for non-existent services. As the number of job postings continues to surge, distinguishing legitimate opportunities from fraudulent ones has become increasingly challenging, making the detection of fake job posts a critical concern for both individuals and organizations.

To address this issue, there has been increasing interest in applying machine learning techniques to automatically identify fraudulent job postings. Traditional methods of manual review are time-consuming, inefficient, and prone to errors due to the vast volume of job posts. Consequently, machine learning-based approaches are seen as an effective solution for automating the classification of job postings as either legitimate or fake, providing a scalable and efficient way to combat employment scams.

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

This paper proposes the use of the Random Forest Classifier, a powerful and widely used ensemble learning algorithm, to classify job postings as either real or fraudulent. The Random Forest Classifier is particularly suitable for this task due to its high accuracy, robustness, and ability to handle complex data with numerous features. The model is trained on the Employment Scam Aegean Dataset (EMSCAD), which contains 18,000 labeled job postings, allowing the classifier to learn patterns associated with fraudulent activity. The dataset is comprehensive, encompassing a variety of job scams, which helps in building a model capable of detecting diverse types of fraudulent posts.

The results of the study demonstrate that the Random Forest Classifier is highly effective in predicting the authenticity of job postings, achieving an accuracy of approximately 98%. This high level of performance indicates the potential of machine learning models to address the growing challenge of employment fraud. By providing an automated and reliable tool for identifying fake job posts, this approach can benefit both job seekers and organizations, enhancing trust and reducing the risks associated with fraudulent job opportunities.

II. LITERATURE SURVEY

Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset
AUTHORS: S. Vidros, C. Kolias , G. Kambourakis ,and L. Akoglu,

The critical process of hiring has relatively recently been ported to the cloud. Specifically, the automated systems responsible for completing the recruitment of new employees in an online fashion, aim to make the hiring process more immediate, accurate and cost-efficient. However, the online exposure of such traditional business procedures has introduced new points of failure that may lead to privacy loss for applicants and harm the reputation of organizations. So far, the most common case of Online Recruitment Frauds (ORF), is employment scam. Unlike relevant online fraud problems, the tackling of ORF has not yet received the proper attention, remaining largely unexplored until now. Responding to this need, the work at hand defines and describes the characteristics of this severe and timely novel cyber security research topic. At the same time, it contributes and evaluates the first to our knowledge publicly available dataset of 17,880 annotated job ads, retrieved from the use of a real-life system.

1) An Intelligent Model for Online Recruitment Fraud Detection

AUTHORS: B. Alghamdi, F. Alharby

This study research attempts to prohibit privacy and loss of money for individuals and organization by creating a reliable model which can detect the fraud exposure in the online recruitment environments. This research presents a major contribution represented in a reliable detection model using ensemble approach based on Random forest classifier to detect Online Recruitment Fraud (ORF). The detection of Online Recruitment Fraud is characterized by other types of electronic fraud detection by its modern and the scarcity of studies on this concept. The researcher proposed the detection model to achieve the objectives of this study. For feature selection, support vector machine method is used and for classification and detection,

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

ensemble classifier using Random Forest is employed. A freely available dataset called Employment Scam Aegean Dataset (EMSCAD) is used to apply the model. Pre-processing step had been applied before the selection and classification adoptions. The results showed an obtained accuracy of 97.41%. Further, the findings presented the main features and important factors in detection purpose include having a company profile feature, having a company logo feature and an industry feature.

job dataset. In addition, we proposed a simple and effective ensemble model combining different deep neural network models. Our experimental results illustrated that our proposed ensemble model achieved the highest result with an F1-score of 72.71%. Moreover, we analyze these experimental results to have insights about this problem to find better solutions in the future.

III. EXISTING SYSTEM

In Existing logistic regression , One R classifier performed well when they balanced the dataset and experimented on that. They tried in their work to find out the problems in ORF model (Online Recruitment Fraud) and to solve those problems using various dominant classifiers. detect fraud exposure in an online recruitment system. They experimented on EMSCAD dataset using machine learning algorithm. They worked on this dataset in three steps- data pre-processing, feature selection and fraud detection using classifier. In the preprocessing step, they removed noise and html tags from the data so that the general text pattern remained preserved. They applied feature selection technique to reduce the number of attributes effectively and efficiently.

Advantages of the Existing System:

Improved Clustering Accuracy: Uses EM to refine clusters with labeled and unlabelled data.
Adaptive Distance Metric: Adjusts distance function for better cluster separability.
Effective Semi-Supervised Learning: Leverages limited labeled data efficiently.
Handles High-Dimensional Data: Enhances feature selection and reduces dimensionality.
Robust to Noise & Outliers: Dynamically adjusts cluster boundaries for stability.
Scalable & Convergent: EM ensures stable results and scales to large datasets.

Disadvantages of existing system:

High Computational Cost: EMSCAD simulations are resource-intensive, which can increase training time and expenses.
Limited Data: EMSCAD may not generate enough diverse data for robust ML models.
Interpretability: ML models may be difficult to interpret, especially in critical design applications.
Integration Issues: Combining EMSCAD with ML tools may require custom solutions and lead to compatibility challenges.
Overfitting: Lack of diverse data can cause ML models to overfit, reducing generalization.
Feature Engineering: Identifying the right features for ML models can be complex without domain knowledge.

ISSN: 2582 - 6379
Orange Publications
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2025, Volume 6 Issue 1
www.ijisea.org

Scalability: EMSCAD simulations may not scale well for large datasets or real-time applications.

V.PROPOSED SYSTEM

The proposed system leverages the Random Forest algorithm, an ensemble learning technique, to improve predictive accuracy and robustness. By constructing multiple decision trees and aggregating their outputs, the system enhances performance while reducing overfitting.

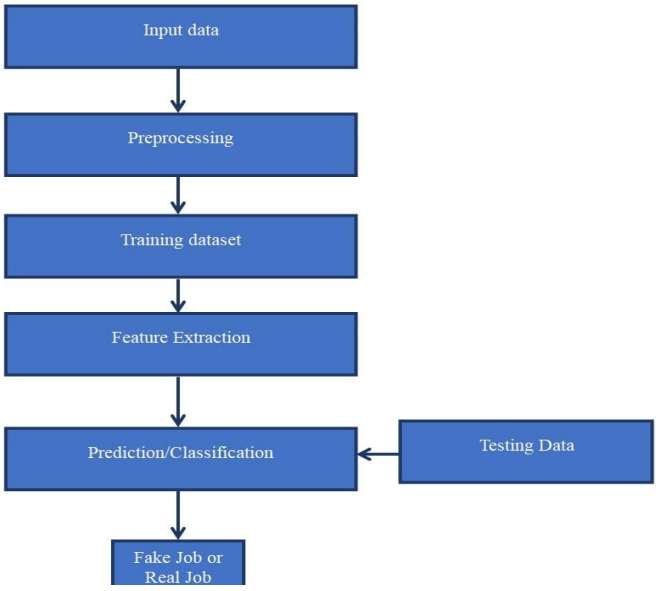
This approach is particularly effective for classification and regression tasks, offering high accuracy, scalability, and interpretability. The system will preprocess input data, train multiple decision trees on random feature subsets, and combine their predictions through majority voting (for classification) or averaging (for regression).

The implementation will be optimized for speed, efficiency, and real-world applicability, making it suitable for domains such as healthcare, finance, and fraud detection.

Key Features of the Proposed Random Forest System

Ensemble Learning Approach – Combines multiple decision trees to improve accuracy and stability.
High Predictive Performance – Reduces overfitting and variance through bagging and random feature selection.
Handles Missing Data – Can process incomplete datasets effectively without significant performance loss.
Scalability & Efficiency – Suitable for large datasets with parallel processing capability.
Feature Importance Ranking – Identifies the most influential features in predictions.
Versatile for Classification & Regression – Works well across different problem domains.
Robust to Noise & Outliers – Less sensitive to noisy data compared to single decision trees.
Low Parameter Tuning Requirement – Works well with default settings while allowing fine-tuning for optimization.

ARCHITECTURE/DATA FLOW DIAGRAM



Advantages of the Proposed System:

1. This approach reduces the number of trainable attributes effectively with less processing time.
2. We have achieved approximately 98% classification accuracy (highest) for Random Forest classifier.
3. We have analyzed performance analysis parameters also to check if the model works well at both false positive and false negative samples.
4. Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, Random Forest Classifier has proved the best prediction results.

Modules of the Random Forest System:

1. **Data Preprocessing** – Cleans, normalizes, and splits data.
2. **Bootstrap Sampling** – Selects random subsets for training.
3. **Decision Tree Construction** – Builds multiple trees with different features.
4. **Prediction Aggregation** – Combines tree outputs (voting/averaging).
5. **Model Evaluation** – Assesses performance with accuracy, RMSE, etc.
6. **Feature Importance** – Identifies key influencing features.
7. **Deployment & Inference** – Deploys model for real-time predictions.

IV. CONCLUSION

Job scam detection has become a great concern all over the world at present. In this paper, we have analyzed the impacts of job scam which can be a very prosperous area in research filed creating a lot of challenges to detect fraudulent job posts. We have experimented with EMSCAD dataset which contains real life fake job posts. In this paper we have experimented both machine learning algorithms (SVM, KNN, Naive Bayes, Random Forest and MLP) and deep learning model (Deep Neural Network). This work shows a comparative study on the evaluation of traditional machine learning and deep learning-based classifiers. We have found highest classification accuracy for Random Forest Classifier among traditional machine learning algorithms and 99 % accuracy for DNN (fold 9) and 97.7% classification accuracy on average for Deep Neural Network.

REFERENCES:

[1] 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.

[2] 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 .

[3] 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), 2020.

[4] Jiawei Zhang, Bowen Dong, Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection witDeep Diffusive Neural Network", IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[5] Scanlon, J.R. and Gerber, M.S., "Automatic Detection of Cyber Recruitment by Violent Extremists", Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s13388-014-0005-5

[6] Y. Kim, "Convolutional neural networks for sentence classification," arXivPrepr. arXiv1408.5882, 2014.

[7] 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-GRU-LSTM-CNN Model," arXivPrepr. arXiv1911.03644, 2019.

[8] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing, vol. 174, pp. 806814, 2016.

[9] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.

[10] K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts," International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205-1209.