A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques

Mr.B.Prasad[1], Janvi Agarwal[2], Satti Shashank Kumar Reddy[2], B Badri Vishal
Reddy[2] Bode Chandra Sekhar[2][1]*Associate Professor,Computer Science and*
Engineering, CMR Engineering College, medchal, T.S, India[2]*B.Tech, Computer*
Science and Engineering, CMR Engineering College, medchal, T.S, India

Abstract: With the development of social media and contemporary technologies, advertising new job openings has recently become a very prevalent problem in the current world. Therefore, everyone will have a lot of reason to be concerned about bogus job postings. Fake job posing prediction presents a variety of difficulties, much as many other categorization problems. In order to determine if a job posting is legitimate or fake, this study proposes using several data mining methods and classification algorithms such KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perceptron, and deep neural network. 18000 samples from the Employment Scam Aegean Dataset (EMSCAD) were used in our experiments. For this classification challenge, a deep neural network classifier excels. For this deep neural network classifier, three thick layers were employed. A bogus job advertisement may be predicted with a classification accuracy of around 98% by the trained classifier using DNN.

1. INTRODUCTION

The advancement of business and technology in the current day has given job searchers a great deal of chance to find new and varied positions. The adverts for these job openings enable job searchers choose their alternatives based on their availability, qualifications, experience, appropriateness, etc. The effect of social media and the internet on the hiring process has increased. The effectiveness of a recruiting process depends on how well it is advertised, therefore social media has a significant influence here. Job information may now be shared in ever-new ways thanks to social media and electronic media marketing. Instead of this, the possibility to disseminate job advertisements quickly has increased the number of fraudulent job postings, which irritate job searchers. People don't respond to fresh job advertisements because they want to keep their personal, academic, and professional information secure and consistent. The genuine goal of legitimate job advertisements through social and electronic media therefore has a very difficult struggle to win over people's trust and trustworthiness[1]. Technologies are all around us to improve and ease our lives, not to create unsafe working conditions. Recruiting new personnel will improve greatly if job postings can be correctly screened to identify fake job postings. False job postings make it difficult for job seekers to locate the positions they desire, which is a significant waste of their time. A new door is opened to deal with challenges in the area of human resource management by an automated system that predicts fake job postings.

A. Job Scam: Fake Job Posting

The term "job scam" refers to online job advertising that are false and often eager to steal the personal and professional information of job searchers in lieu of providing them with suitable positions. Fraudulent individuals sometimes attempt to steal money from job seekers. According to a recent poll by Action Fraud in the UK, more than 67% of individuals who hunt for employment online but are ignorant of fraudulent job postings or job scams are at significant risk [2]. Nearly 700,000 job searchers in the UK reported losing more than $500 000 as a result of work scams. The survey indicated an almost 300% growth in the UK during the previous two years. Students and recent graduates are the main targets of fraudsters since they often want to get a stable job for which they are prepared to spend more money. Techniques for avoiding or protecting against cybercrime fall short because con artists regularly alter their methods of employment fraud.

B. Typical forms of job scams

Fraudsters produce bogus job adverts in order to get other people's personal information, such as insurance data, bank details, income tax details, date of birth, and national id. Advance fee scams happen when con artists demand money while using justifications such as administrative fees, information security testing costs, management costs, etc. Sometimes con artists pose as employers and inquire about applicants' passport information, bank account information, driving records, etc. as a pre-employment screening.

A Description of The Project:

When they get students to deposit money into their accounts and subsequently transfer it back, money laundering frauds take place [3]. This "cash in hand" strategy results in work with cash on hand without having to pay any taxes. In order to lure job searchers, scammers often develop bogus corporate websites, bank websites, official-looking papers, etc. Instead of engaging in face-to-face conversation, the majority of employment fraudsters attempt to capture victims through email. To establish themselves as headhunters or recruiting agency, they often use social networking sites like LinkedIn. They often work to provide the job seeker the most accurate representation of their business profile or websites. Regardless of the employment scam they use, they constantly try to lure job seekers into their traps by gathering information and using it to their advantage to either generate money or accomplish other goals.[4].

2. LITERATURE SURVEY

2.1 Existing System

To determine if a job posting is authentic or false, several studies have been conducted. A significant amount of study is being done to identify employment fraud online. Job fraudsters were referred to be phoney online job advertisers by Vidros et al [5]. They discovered data regarding several legitimate and well-known businesses and organisations that created false job adverts or vacancy listings with ulterior motives. On the EMSCAD dataset, they conducted experiments employing a variety of classification techniques, including the naïve bayes classifier, random forest classifier, Zero R, and One R. The dataset's highest performance was shown by the Random Forest Classifier, which had a classification accuracy of 89.5%. They discovered that the dataset had relatively low logistic regression performance. When the dataset was balanced and tested, one R classifier performed well. They made an effort in their research to identify the issues with the ORF model (Online Recruitment Fraud) and to address those issues utilising other dominant classifiers [6].

A methodology to identify fraud exposure in an online recruiting system was put out by Alghamdi et al. On the EMSCAD dataset, they conducted experiments using a machine learning method. They worked on this dataset in three stages: feature selection, data pre-processing, and classifier-based fraud detection. In order to retain the overall text pattern, they deleted noise and html tags from the data during the preparation stage. To effectively and efficiently limit the amount of characteristics, they used the feature selection approach. Support Vector Machine was used to determine the features, and a random forest ensemble classifier was utilised to identify bogus job postings from the test data. With the aid of the majority voting approach, the random forest classifier seemed to be a tree-structured classifier that operated as an ensemble classifier. With 97.4% classification accuracy, this classifier was able to identify bogus job postings.

Different deep neural network models, such as Text CNN, Bi-GRU-LSTM CNN, and Bi-GRU CNN, which are pre-trained using text datasets, have been suggested by Huynh et al. They sought to categorise the dataset of IT jobs. They trained a TextCNN model with a convolution layer, a pooling layer, and a fully connected layer using data from IT jobs. This model used convolution and pooling layers to train data. The weights were flattened and then transferred to the layer with all connections. This model's classification method employed the softmax function. To improve classification accuracy, they also utilised an ensemble classifier (Bi-GRU CNN, Bi-GRULSTM CNN) utilising a majority voting approach. They discovered that TextCNN had a classification accuracy of 66% and Bi-GRU-LSTM CNN had a classification accuracy of 70%. The ensemble classifier, which had an accuracy of 72.4%, completed the classification job the best.

In order to differentiate between real and false news (including articles, creators, and topics) using text processing, Zhang et al [7]. suggested an automated fake detector model. They have employed a unique dataset of news or items shared on Twitter via the PolitiFact website account. To train the suggested GDU diffusive unit model, this dataset was used. This trained model performed well as an automated fake detecting model when input came from numerous sources at once.

2.2Proposed System

To identify bogus job postings, the system has employed EMSCAD. Each row of the data in this dataset has 18 characteristics, including the class label, and there are 18000 samples in total. The characteristics include employment_type, required_experience, required_education, industry, function, fraudulent (class label), job_id, title, location, department, salary_range, company_profile, description, requirements, benefits, and telecommunication. Only 7 of these 18 traits, which are transformed into category attributes, have been utilised. Work from home, has a firm logo, has questions, job type, necessary education, required experience, and fraudulent are converted from text values to category values. As an example, the values for "employment_type" are changed to 0 for "none," 1 for "full-time," 2 for "part-time," 3 for "others," 4 for "contract," and 5 for "temporary." The major reason for converting these characteristics into categories is to categorise false job postings without using text processing or natural language processing. We have solely utilised those category characteristics in this study.

2.2 Preliminary cryptographic steps

Dataset

EMSCAD has been used to identify bogus job postings. Each row of the data in this dataset has 18 characteristics, including the class label, and there are 18000 samples in total. The characteristics include employment_type, required_experience, required_education, industry, function, fraudulent (class label), job_id, title, location, department, salary_range, company_profile, description, requirements, benefits, and telecommunication. Only 7 of these 18 traits, which are transformed into category attributes, have been utilised.

Telecommuting, has_company_logo, has_questions, employment_type, necessary education, required experience, and fraudulent are converted from text value to categorical value. For instance, the values for "employment_type" are changed to 0 for "none," 1 for "full-time," 2 for "part-time," 3 for "others," 4 for "contract," and 5 for "temporary." The major reason for converting these characteristics into categories is to categorise false job postings without using text processing or natural language processing. We have solely utilised those category characteristics in this study.

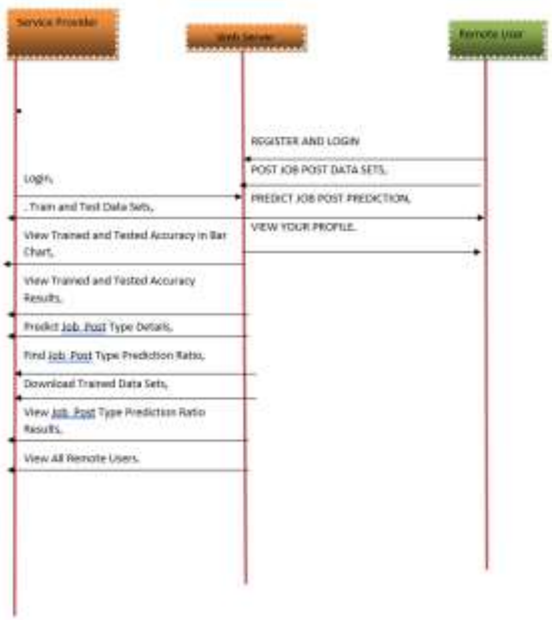
Service Provider
Web Server
Remote User
REGISTER AND LOGIN
POST JOB POST DATA SETS,
PREDICT JOB POST PREDICTION,
VIEW YOUR PROFILE.
Login,
.Train and Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
Predict Job Post Type Details,
Find Job Post Type Prediction Ratio,
Download Trained Data Sets,
View Job Post Type Prediction Ratio Results,
View All Remote Users.

Fig 1: System Flow

3. SYSTEM ANALYSIS&DESIGN

3.1 Neural Network

Neural networks operate on the fundamental tenets of how the human brain functions. It enables a computer to determine how much two patterns resemble or vary from one another by comparing them. A neuron is a mathematical function that extracts characteristics and categorises certain patterns [8].

There are several layers of interconnected nodes in a neural network. Each perceptron node functions as a multiple linear regression. The result of multiple linear regression is passed through this perceptron and converted into a non-linear activation function. Perceptrons are organised in layers that are linked to one another[9]. To reduce mistake rates, the hidden layers change the weights of the input layers. The neural network functions as a classifier for supervised learning[10].

Deep neural networks, part B

Deep neural networks (DNNs) are Artificial Neural Networks (ANNs) that include several layers between the input and output layers. The feed forward algorithm powers DNN. From the input layer to the output layer, data flow is directed [11]. DNN generates a large number of virtual neurons that have their connection weights initialised with random numbers. This weight is multiplied by the input, and the result is an output that ranges from 0 to 1. To effectively categorise the output, the weights are adjusted throughout training.

The model overfits as a result of learning unusual patterns from additional layers. Dropout layers allow for a generalisation of the model by reducing the number of trainable parameters. In this study, we utilisedrelu as the activation function and adam as the optimizer to train a sequential model with dense layers on the data. Adam computes individual learning rates based on many factors during the training process since this is an adaptive learning approach [12].

Additional classifiers

Our work dataset is trained on the classifiers K Nearest Neighbour, Random Forest Classifier, Decision Tree, Naive Bayes Classifier, Support Vector Machine (RBF kernel), and Multilayer Perceptron (MLP).

3.3 System Architecture:

Precision equals TP+TN/TP+FP+FN+TN.

Precision equals TP/TP+FP

F1 Score = 2*(Recall * Precision) / (Recall + Precision) (TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative)

In order to train the data for the deep neural network model, 10 fold cross validation is employed. A total of 60% of the data was utilised for training, 20% for determining validation accuracy, and the remaining 20% for testing the model's effectiveness. The validity accuracy reveals the model's degree of performance with respect to unobserved data [13].

In each training period, we have seen a positive correlation between validation and training accuracy. We may identify the trained model as a generalised one if the validation accuracy is greater than the training accuracy. We utilised a dropout layer to lessen the model's overfitting. In order for the model to function effectively outside of the training dataset, this layer decreases the trainable parameters at each round of training [14].

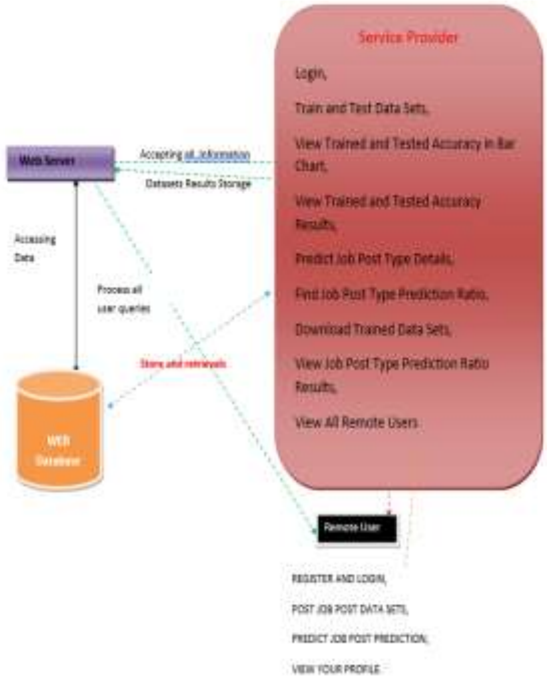
Service Provider
Login,
Train and Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
Predict Job Post Type Details,
Find Job Post Type Prediction Ratio,
Download Trained Data Sets,
View Job Post Type Prediction Ratio Results,
View All Remote Users
Web Server
Accepting all information
Datasets Results Storage
Accessing Data
Process all user queries
Store and retrievals
WEB Database
Remote User
REGISTER AND LOGIN,
POST JOB POST DATA SETS,
PREDICT JOB POST PREDICTION,
VIEW YOUR PROFILE

Fig 2 : System Architecture

3.4 Data Flow Diagram : Whenever a new system is developed, user training is required to educate them about the working of the system so that it can be put to efficient use by those for whom the system has been primarily designed [15]. For this purpose the normal working of the project was demonstrated to the prospective users. Its working is easily understandable and since the expected users are people who have good knowledge of computers, the use of this system is very easy.

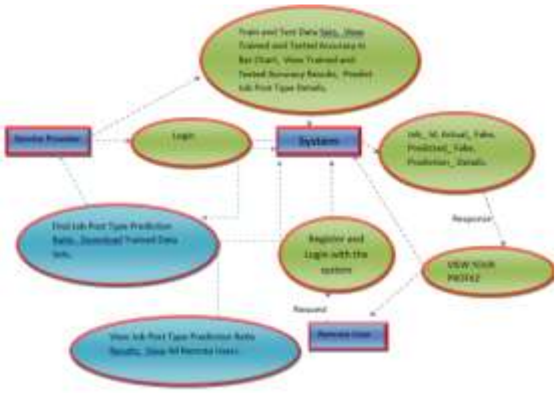
Train and Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, Predict Job Post Type Details,
Service Provider
Login
System
Job_Id, Actual_ Fake, Predicted_ Fake, Prediction_ Details
Find Job Post Type Prediction Ratio, Download Trained Data Sets,
Register and Login with the system
Response
VIEW YOUR PROFILE
Request
View Job Post Type Prediction Ratio Results, View All Remote Users
Remote User

Fig 3: Data Flow Diagram

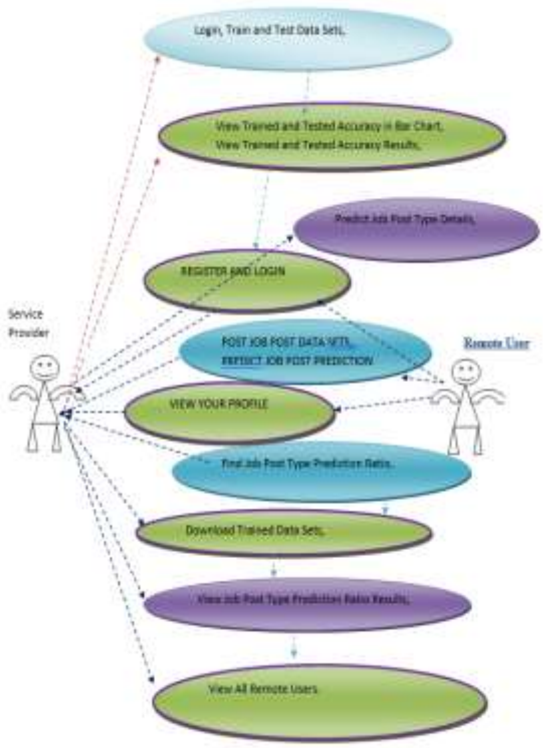
Login, Train and Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
Predict Job Post Type Details,
REGISTER AND LOGIN
Service Provider
POST JOB POST DATA SETS,
PREDICT JOB POST PREDICTION
Remote User
VIEW YOUR PROFILE
Find Job Post Type Prediction Ratio,
Download Trained Data Sets,
View Job Post Type Prediction Ratio Results,
View All Remote Users

Fig 4 Use Case UML Diagrams

5. CONCLUSION

The identification of job scams has recently become a major problem worldwide. We have examined the effects of employment scams in this article since they might be a highly lucrative topic of study and make it difficult to identify fake job postings. We conducted experiments using the EMSCAD dataset, which comprises real-world fictitious job postings. In this research, we experiment with both deep learning (Deep Neural Network) and machine learning (SVM, KNN, Naive Bayes, Random Forest, and MLP). This article presents a comparison study on the assessment of classifiers based on deep learning and conventional machine learning.

In comparison to other conventional machine learning methods, Random Forest Classifier has the greatest classification accuracy. DNN (fold 9) and Deep Neural Network have the highest classification accuracy on average.

6. FUTURE SCOPE

This article presents a wide range of machine learning techniques for detecting employment fraud with the ultimate goal of addressing the issue of fraud in the workplace. In the future, a supervised technique will be used to illustrate how different classifiers may be used to the detection of employment fraud. The results of the studies show that the Random Forest classifier outperforms the rivals' classifiers in terms of accuracy. The literature supports this assertion. The recommended approach has an accuracy rate of 98.27 percent, which is much higher than the accuracy rate of the existing methods when compared to current procedures.

7. 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] Santosh Kumar Srivastava Dr. Yogesh Kumar Sharma Dr. Yogesh Kumar SharmaSheo Kumar, " Precision enhancement of Intrusion detection system through outlier detection and feature classification",

International Journal of Control and AutomationVol. 12, No. 6, (2019), pp. 820-830, ISSN: 2005-4297 IJCA.

[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] Shikha Agrawal and Rajesh Tiwari, “Enhancing and Performance Comparison of various Truth Discovery Approach”, International Journal of Technology, Vol. 1, Issue 2, July – December 2011, pp 76–86, ISSN(online): 2231- 3915 ISSN(print): 2231- 3907.

[5] Scanlon, J.R. and Gerber, M.S., “Automatic Detection of Cyber Recruitment by Violent Extremists”, Security Informatics, 3, 5, 2014,

[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. 806 814, 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] Santosh Kumar Sriavastava1, Dr. Yogesh Kumar Sharma2, Dr. Sheo Kumar3, “Using Of WEKA Tool In Machine Learning: A Review" ,

International Journal of Advanced Science and TechnologyVol. 29, No. 6, (2020), pp. 8604-8614`8604ISSN: 2005-4238 IJASTCopyright ⓒ 2020.

[11] Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security& Its Applications, 8, 55-72.

[12]] Rajesh Tiwari, Manisha Sharma, Kamal K. Mehta and Mohan Awasthy, “Dynamic Load Distribution to Improve Speedup of Multi-core System using MPI with Virtualization”, International Journal of Advanced Science and Technology, Vol. 29, Issue 12s, 2020, pp 931 – 940, ISSN: 2005 – 4238.

[13] Rajesh Tiwari, Manisha Sharma and Kamal K. Mehta, "Performance Analysis of various High-Performance Computing Models" , Journal of Advanced Research in Dynamic and Control System, Vol. 10, special issue 2, July 2018, pp 2192 – 2200 , ISSN: 1943 – 023X

[14]Dr. C.N. RAVi , D. Palanivel Rajan, Desa Uma Vishweshwar, Edem Sureshbabu (CMREC)'A Review on Various Cloud-Based Electronic Health Record Maintenance System for COVID-19 Patients'Name of the Journal with ISSN:*Advances in Cognitive Science and Communications*, Cognitive Science and Technology - Springer Nature Singapore Pte Ltd. 2023 (CMREC Conference- ICCCE-2022)Vol. / Issue /PP. No. / Date/Month & Year of Publication:Impact Factor: https://doi.org/10.1007/978-981-19-8086-2_15, April 2023.

[15] Mrs.G.Sumalatha1 , Y.Jaideep Naidu M.Srividya, Karra karthik Reddy , D.Niharika, 'Smart OCR for Document Digitization', JASC: Journal of Applied Science and Computations, ISSN NO: 1076-5131, Volume VIII, Issue III, Macrh/2021.