RESEARCH PAPER

on

A COMPARATIVESTUDYON FAKE JOB POSTING PREDICTION USING ML TECHNIQUES

(CSE VIII Semester Major project 2022-23)

Submitted by :

Pooja Nauni, Roll No.: 1918537/47

CSE-E-VIII-Sem Session: 2022-23

Submitted to:

Miss. Preeti Chaudhary
Asst. Professor
(Department of CSE)

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

GRAPHIC ERA HILL UNIVERSITY, DEHRADUN

A COMPARATIVESTUDYON FAKE JOB POSTING PREDICTION USING ML TECHNIQUES

ABSTRACT

In the modern culture of new trends and technologies Internet have a great impact on the recruitment process and job posting methods. It became very easy to post a job vacancy for job recruiters on social media and electronic media. So, Employment Spam Detection has become an area of great concern to all. Like many other classification tasks, fake job posting prediction leaves a lot of challenges to face. In this project we have used the popularly known classification algorithm Random Forest to predict a job post if it is real or fraudulent. The project is the experimental data analysis of the EMSCAD dataset. The classifier trained using this algorithm shows the classification accuracy of approximately 97% to predict a fake job post.

INTRODUCTION

In the modern era, where the world's population is increasing day by day the number of job seekers is also increasing exponentially, especially in a country like India, China, etc. The vast advancements in development and technology have opened many new opportunities for youth who are unemployed. The very common media to reach out to the public who wants to work is through advertisements on electronic and social media. Advertisements on electronic and social media have created newer and much easier ways for job providers to share job postings and their details to reach out to large and efficient people for the job. The whole recruitment/hiring process is influenced by the internet and there is the tremendous impact of social media on it. Earlier, the average age to get a job was around 25-29 years but nowadays much younger people are getting hired at the age of 21 or 22.

With increasing ways to share job postings the percentage of fake job postings has also increased which leads to harassment for the job seekers. It creates a feeling of insecurity in their minds. An efficient solution to this problem can be a system that can classify job postings into two categories fake or real. This system will be a great advancement in the field of Human Resource Management, and it will build confidence in people to go ahead with the jobs which are posted on electronic media or social media.

History of the topic

There are many methods that can be helpful to predict whether a job post is fake or real. In this project, the methodology we have used is classification techniques to classify the jobs into two classes: fake and real. We have trained the model emscad dataset using the Random Forest classifier for the bifurcation of jobs using their Job Descriptions and some details about the job posts. This trained classifier acts as an online fraud job posting detector.

Dataset:

The dataset is the emscad dataset which we have approx.(18000) samples of the jobs and each row of the dataset contains attributes/columns in it which helps to detect whether a job post is fake or real. It contains both the meta – information as well as textual information about the jobs.

This dataset is very useful as based on these following things can be done:

- Classification model can be developed based on textual and meta information to detect fraudulent jobs
- The key features of fake or real jobs can be identified
- Data Analysis can be performed to identify interesting insights from thisdataset.

Technology involved

The technology that we have used in this project is as follows :

- **Python:** Python is an interpreted, high-level general purpose and object-oriented programming language. It is a language that lets developers work quickly and integrate systems more effectively. It is widely used for Machine Learning.

- **Supervised Machine Learning:** Supervised Machine learning is commonly used to classify the data samples or make predictions accurately. It uses labeled data for training the model.

- **Natural Language Processing:** Natural language Processing is the component of Artificial Intelligence that has the capability of a system to understand human language as it is spoken.

Motivation

The motivation behind doing this project is to save the time of youth by one click detection of any job description and prevent youth from getting harassed by these fraudulent job posts. There are many cases registered into the cybercrime department where using these fake job posts young people are blackmailed by stealing their confidential information and many people are asked to deposit money to get their jobs and many other serious crimes are committed using these fraudulent job postings. Apart from this the main agenda of doing this project is to learn and understand various machine learning algorithms which are used to develop such kind of classifiers.

Problem Statement

Currently, various fraudulent recruiters are regularly posting fake job posts on their social media handles to make people believe on them and then steal their personal and confidential information using which people are getting blackmailed. A recent survey done by ActionFraud from UK has clearly mentioned in their survey report that more than 67% people are at great risk who seek jobs through online advertisements but are not aware of fake job posts or job scams happening all around the globe. There are various consequences of such fraudulent activities as people believing in them can unintentionally become a part of serious criminal activities without their knowledge. Mostly young people who are students or fresh graduates are targeted because they do have the pressure of getting a job and securing their lives as soon as possible. We seek to resolve this issue by doing a deep study in this field.

Objective

The aim of this project is to develop an automated system that acts as a classifier that will have the capability to identify whether the job posting is fake or real. It takes textual or numeric data as input and based on that classifies the test data into real or fake. It uses various parameters and features to judge whether it is real or fraudulent.

LITERATURE SURVEY

In recent years, online scam is a major problem that has affected a lot of people's lives. To reduce this many preventions were taken like fake news detection, Email spam detection, and Review spam detection. These all fall in the category of online fraud detection.

Author	Published year	Technology used	dataset involved	Description
Priya Khandagale, Akshata Utekar, Anushka Dhonde, Prof. S. S. Karve	2022	Machine Learning: Supervised Learning and Random Forest Classifier	'Pandas_datareader'	With the help of this project, j seekers will get only those jobs that is legal. Supervised the mechanism is used to exemplify the use of several classifiers for employee sca detection.
Syed Ishfaq Manzo Dr Jimmy Singla, Nikita	2019	Deep Diffusive Neural Network	LIAR	They have classified every tweet as binary classification tweets. They have got an accuracy between 63 and 70 percentage.
Bhandavya K R, Dr M.N Veena	2022	Machine Learning	Employment Scam Aegean Dataset	With the help of the Random Forest classifier (Machine Learning), they were able to differentiate between honest and dishonest jobs alert.
Shawni Dut Prof. Sameer Kumar Bandyopadhyay	2020	Machine Learning	Dataset is taken from "Kaggle"	According to them, the Random Forest classifier is the best method to detect fake jobs.
Sultana Umne Habiba, Md. Khairul Islam, Farzana Tasmin	2021	Data Mining techniques and Machine Learning classification algorithm	Employment Scam Aegean Dataset	With this dataset's help, they could successfully detect fake job posts. At first, they removed all the noise and HTML tags, then the number of attributes, and finally with the help of a classifier they were to detect most votes for fake posts.
Marcel Naud Rohan Nand Kolawole John Adebayo	2022	Machine Learning	Employment Scam Aegean Dataset	They have divided the fake job into three categories: identity theft, corporate identity theft, and pyramid schemes or multi-level marketing
Vidros, Koli Kambourakis an Akoglu	7	Machine Learning	Employment Scam Aegean Dataset	Here they compared a model of a bag of words and a handcrafted binary rule-set and they achieved an accuracy of 91%.
B. Alghamdi, Alharby	9	Machine Learning	Employment Scam Aegean Dataset	He proposed a model with the help EMSCAD dataset. They achieved an accuracy of 97.4% wi the help of the Random Fore Classifier
Tin Van Huynh, Kiet Van Nguy Ngan Luu-Th Nguyen1, and A Gia-Tuan Nguyen	2020	Deep Diffusive Neural Network	IT job dataset	With the help of the IT job dataset, they were able to detect fraud job posting. Majority of votes using ensemble classifier, were useful for predicting true jobs.
Jiawei Zhan Bowen Dong, Phi S. Yu,	0	Deep Diffusive Neural Network	The custom dataset used by Politifact website Twitter accoun	With the help of text processing, they were able to identify the fa and true job posts. The propos dataset were used to modularize the GDU diffusive unit.

METHODOLOGY USED

The methodology that we have used in this project is a Supervised Machine Learning algorithm. We have used the Random Forest classification algorithm to classify the data into fake or real. Basically, the work of this project is to identify the key features using which the test data can be classified whether its fake or real.

Data Preprocessing: We started the preprocessing of dataset by creating the lists of stop words and punctuation marks and load the English tokenizers, staggers, parsers and words. After that we created the tokenizer function and object and then convert each token into lowercase and remove all the stop words, punctuation marks and white spaces.

Next we have used TF-IDF(term frequency-inverse document frequency) vectorizer that gives equal weightage to all the words and also shows how important the word is.

After the preprocessing of dataset we split the dataset into test and train with the split of 30% and 70% respectively, and then applied the random forest classifier.

Random Forest Classifier: It creates a set of decision trees from a randomly selected subset of our training dataset. It is a set of decision trees from a randomly selected dataset.

The algorithm given below explains the working of the Random Forest Algorithm on a dataset:

Step 1: Select random samples from the given data or training set.

a

Step 2: This algorithm will construct a decision tree for every training dataset.

Step 3: By calculating the average of the decision trees, we can find its overall vote.

Step 4: Finally, select the most voted prediction result as the result for the prediction. **Mathematical**

formula involved:

To construct a decision tree, our classifier will be going to use this formula.

Entropy: It is the measurement of unpredictability or impurity in the system.

Information Gain: It measures the reduction in entropy and decides which attribute should be selected as the decision node.

Entropy(s)=-P(yes)log$_2$P(yes)- P(no) log$_2$ P(no) Where,

S is the total sample space

P(yes) isprobability of yes

P(no) is probability of no

If numberof yes= number of no i.e. P(S)=0.5

Entropy(s)=1

If numberof yes or all no i.e. P(S)=1 or 0

Entropy(s)=0

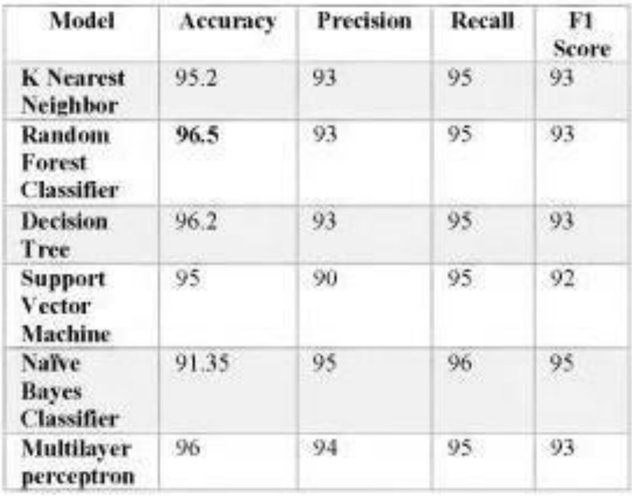
If S is our total collection,

Information Gain=Entropy(S)-[(Weighted Avg)* Entropy(each feature)

RESULTS

We worked on the dataset in three steps data pre-processing, feature selection, and fraud detection using the classifier. Random forest classifier will work as ensemble classifier with the help of majorityvoting technique. Using this classifier, we achieved approximately 97% accuracy.

Based on our literature survey we concluded that applying different algorithms on EMSCAD dataset, it gave the below result.



Model	Accuracy	Precision	Recall	F1 Score
K Nearest Neighbor	95.2	93	95	93
Random Forest Classifier	**96.5**	93	95	93
Decision Tree	96.2	93	95	93
Support Vector Machine	95	90	95	92
Naïve Bayes Classifier	91.35	95	96	95
Multilayer perceptron	96	94	95	93

Table 1: Comparison of Various machine Learning Classifiers

In above Table 1 the Classification Accuracy, Precision, recall and F1 score of all these classifiers are showing highest accuracy using Random Forest Classifier.

So, we have analyzed f1 score to check if the model works well at both false positive and false negative samples.

Using following formulas of accuracy, Precision, f1 score and recall

$$\text{Accuracy} = TP+TN/TP+FP+FN+TN$$

$$\text{Precision} = TP/TP+FP, \text{Recall} = TP/TP+FN,$$

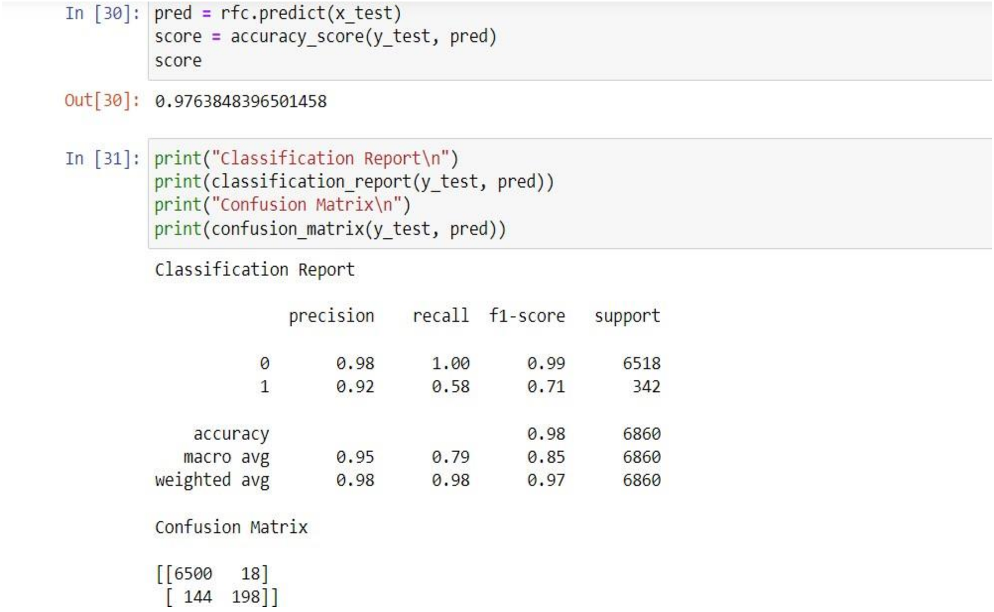
$$\text{f1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

Precision identifies the ratio of correct positive results over the number of positive results predicted by the any classifier.

Recall denotes the number of correct positive results / the number of all relevant samples.

F1-Score is a parameter that is concerned for both recall and precision and it is calculated as the harmonic mean of precision.

We obtained the following accuracy and confusiom matrix:



```
In [30]: pred = rfc.predict(x_test)
         score = accuracy_score(y_test, pred)
         score

Out[30]: 0.9763848396501458

In [31]: print("Classification Report\n")
         print(classification_report(y_test, pred))
         print("Confusion Matrix\n")
         print(confusion_matrix(y_test, pred))

         Classification Report

                       precision    recall  f1-score   support

                    0       0.98      1.00      0.99      6518
                    1       0.92      0.58      0.71       342

             accuracy                           0.98      6860
            macro avg       0.95      0.79      0.85      6860
         weighted avg       0.98      0.98      0.97      6860

         Confusion Matrix

         [[6500   18]
          [ 144  198]]
```

CONCLUSION

At present, Job scam detection has become an area of great concern across the globe. Studying the topic we analyzed the impact of job scams on today’s youth very prosperous research area with a lot of challenges to detect fraudulent job postings. In this project we have trained our model on EMSCAD which is vast enough to develop an efficient model containing real-life job posts some of which are fake, and some are real. In this paper we have compared classifiers of various Machine Learning Algorithms. The highest classification accuracy is detected using Random Forest Classifier among traditional machine learning algorithms which is 97%. This proposal will guide job seekers to get real and legitimate job offers from various job recruiters. Classifiers based on the supervised mechanism are used for Employment Spam Detection.

REFERENCES

1. Sultana Umme Habiba, Md. Khairul Islam, and Farzana Tasnim, “A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques”, 2021 2nd International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST).
2. Bhandavya K R1 , Dr M.N Veena2, “A Comparative Study of Fake Job Post Prediction Using Different Data mining Techniques”, International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified│□Impact Factor 7.39│□Vol. 11, Issue 7, July 2022 DOI: 10.17148/IJARCCE.2022.11751
3. Marcel Naudé1 · Kolawole John Adebayo2,3 · Rohan Nanda4,5, “A machine learning approach to detecting fraudulent job types”, AI & SOCIETY https://doi.org/10.1007/s00146- 022-01469-0.
4. Priya Khandagale1 , Akshata Utekar2 , Anushka Dhonde3 , Prof. S. S. Karve4, “Fake Job Detection Using Machine Learning”, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue IV Apr 2022.
5. Shawni Dutta#1 and Prof.Samir Kumar Bandyopadhyay*2, “ Fake Job Recruitment Detection Using Machine Learning Approach”, International Journal of Engineering Trends and Technology (IJETT) – Volume 68 Issue 4- April 2020.
6. 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.
7. B. Alghamdi, F. Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, 2019, Vol 10, pp. 155 176, https://doi.org/10.4236/iis.2019.103009.
8. 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.
9. 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
10. S. I. Manzoor, J. Singla and Nikita, "Fake News Detection Using Machine Learning approaches: A systematic Review," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 230-234, doi: 10.1109/ICOEI.2019.8862770.
11. https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction
12. https://in.coursera.org/specializations/machine-learning
13. Supervised Machine Learning : Regression and Classification course by Coursera
14. https://in.coursera.org/learn/introduction-to-ai
15. Data Mining Foundations and Techniquesby Coursera
16. Machine Learning Using Python by Maharajan Pradhan and U Dinesh Kumar
17. The Complete References Python by Martin C. Brown.