Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

FAKE JOB DETECTION USING MACHINE LEARNING

Ms. SADHU THIRUSHA[1],Ms.CHEVVURI SOWJANYA[2] , Ms. SHAIK NAZIYA BEGUM [3],Ms.AVULAPATI TEJASWINI[4],Ms. MVL.SUGUNA PRIYA[5], Mrs. MAKKAPATI LAKSHMI PRASANNA

1.BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108 thirushasadhu2@gmail.com

2. BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108

3. BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108

4. BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108

5. BTECH, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108

6. ASSISTANT PROFESSOR, COMPUTER SCIENCE AND ENGINEERING, VIJAYA INSTITUE OF TECHNOLOGY FOR WOMEN, ENIKEPADU, VIJAYAWADA, ANDHRAPRADESH, INDIA-521108 makkapatiprasanna@gmail.com

ABSTRACT

With the development of social media and modern technologies, advertising new job openings has recently become a very prevalent problem in the current world. So, everyone will have a lot of reason to be concerned about bogus job postings. Fake job posing prediction presents a variety of difficulties, just like many other categorization tasks. In order to determine whether a job posting is legitimate or fraudulent, this paper proposed using various data mining techniques and classification algorithms like 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 used. A bogus job advertisement can be predicted with a classification accuracy of about 98% by the trained classifier using DNN.

1 INTRODUCTION

Single Classifier based Prediction:

In order to forecast the unknowable test instances, classifiers are trained. While identifying bogus job postings, the following classifiers are employed.

A)Narrow Bayes Classifier

A supervised classification method called the Naive Bayes classifier makes use of the Bayes Theorem of Conditional Probability. Even though this classifier's probability predictions are off, the judgement it makes in practise is extremely good. In the following situation—when the characteristics are independent or wholly functionally dependent—this classifier achieves a very promising result. The amount of information lost from the class due to the independence assumption is needed to estimate the accuracy of this classifier, not the feature dependencies.

B) Multi-Layer Perceptron Classifier:

By including optimum training settings, multi-layer perceptrons can be used as supervised classification tools. A multilayer perceptron's hidden layer count and the number of nodes in each layer might vary depending on the situation at hand. The network architecture and training data both influence the parameter selection decision.

C) K-nearest Neighbor Classifier:

using K-Nearest Neighbor Classifiers, also referred to as lazy learners, identify items based on the training examples' closeness to the objects in question in the feature space. When defining the class, the classifier takes into account k numbers of items as the closest object. The main difficulty with this categorization method lies in selecting the right value for k.

D) Decision Tree Classifier:

A classifier that uses a tree-like structure is called a Decision Tree (DT). It learns about classification. Each target class is represented by a leaf node of the decision tree (DT), while non-leaf nodes of the DT are utilised as decision nodes to signal specific tests. Any of the branches of that decision node can determine the results of those tests. This tree is traversed beginning at the root and moving upward until a leaf node is reached. It is a method for getting a decision tree's classification results. Spam filtering has used a method called decision tree learning. By using and refining this model, it can be useful for predicting the target depending on various criteria.

Ensemble Approach based Classifiers:

The ensemble approach enables numerous machine learning algorithms to work together to improve the overall system's accuracy. The notion of ensemble learning approach and regression technique are both used by random forest (RF) to solve classification-based challenges. This classifier combines a number of classifiers that resemble trees and are used on various subsamples of the dataset. Each tree casts a vote for the class that best fits the input. Boosting is an effective method for

increasing classification accuracy in which a number of unstable learners are combined into a single learner. The boosting technique applies a classification algorithm to reweighted versions of the training data and selects the classifier sequence with the weighted majority vote. Ada Boost is a solid illustration of a boosting approach that results in enhanced output even when the weak learners' performance is subpar. Boosting algorithms have a good track record of resolving spam filtration issues.Another boosting technique-based classifier that makes use of decision trees is the gradient boosting algorithm. Moreover, it reduces prediction loss.

2. RELEATED WORK

Internet recruiting fraud detection is a relatively new field that has not seen much research. There are some indirect methods to address online recruitment fraud to a limited extent, such as Email Spam filtering, which prevents sending advertising-related emails to users, anti-phishing techniques to identify fake websites, and anti-opinion fraud safeguards to identify the publication of deceptive and misleading fake reviews. According to numerous studies, the detection of review spam, email spam, and fake news has all attracted considerable attention in the field of online fraud detection.

3. PROPOSED WORK AND ARCHITECTURE

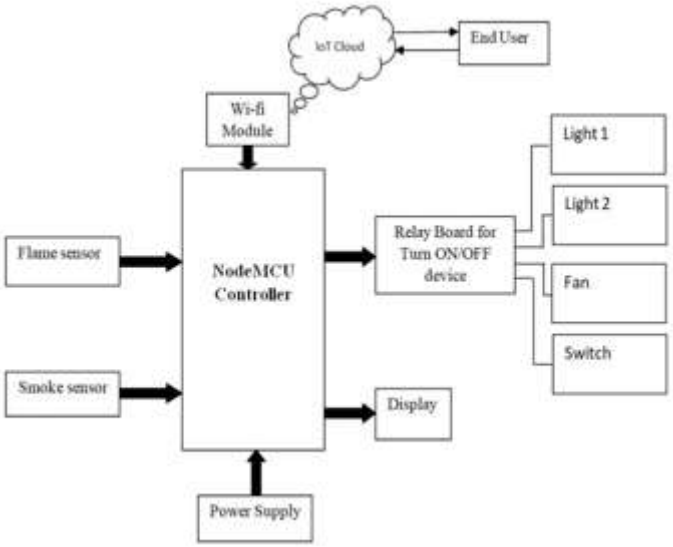
To identify bogus job postings, the system has employed EMSCAD. Each row of the data in this dataset has 18 attributes, including the class label, and there are 18000 samples in total. The characteristics include the job ID, title, department, salary range, company profile, requirements, benefits, telecommunication, corporate logo, questions, employment type, necessary experience, necessary education, industry, function, and bogus (class label). Only 7 of these 18 traits, which are transformed into category attributes, have been used. Telecommuting, has a company logo, has inquiries, is employed, has experience, requires education, and is changed 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 used those categorical attributes in this work.

ADVANTAGES

1) The suggested has been implemented using the quick and precise EMSCAD technique.

2) The system is highly effective since it accurately detects fake job postings, which make it difficult for job seekers to locate the positions they desire and significantly waste their time.

ARCHITECTURE



4. METHODOLOGIES

Supervised Machine Learning algorithm can be broadly classified into Regression and Classification Algorithms.

Regression Algorithms: Regression finds correlations between dependent and independent variables. Regression

algorithms therefore aid in the prediction of continuous variables such as real estate prices, economic trends, climatic patterns, oil and gas prices (a crucial job in today's world!), etc.

Linear Regression

- Decision Tree Regressor
- Random Forest Regressor

Linear Regression:

One of the simplest and most widely used Machine Learning methods is linear regression. It is a mathematical technique for performing predictive analysis. For continuous/real/numeric variables like sales, salary, age, and product price, among others, linear regression generates predictions. The linear regression algorithm, also known as linear regression, demonstrates a linear connection between a dependent (y) and one or more independent (y) variables. Given that linear regression demonstrates a linear connection, it can be used to determine how the dependent variable's value changes as a function of the independent variable's value. The connection between the variables is represented by a sloping straight line in the linear regression model.

Decision Tree Regression:

Decision tree regression can be used to perform non-linear regression in machine learning. The decision tree regression algorithm's primary job is to divide the information into more manageable chunks. The values of all data points that relate to the issue statement are plotted using the subsets of the dataset. This algorithm divides the data collection into decision and leaf nodes, producing a decision tree. When the data collection has not undergone enough change, ML experts favour this model. One should be aware that even a small shift in the data can have a significant impact on the decision tree's structure. Additionally, one should avoid over-pruning the decision tree regressors. since there won't be enough remaining end nodes to make the forecast. One should not overly prune the decision tree regressors in order to have numerous end nodes (regression output values). This develops a model with a tree-like structure that can forecast data in the future and generate useful ongoing output.

Random Forest Regression Algorithm:

Another popular method for non-linear regression in machine learning is random forest. A random forest employs multiple decision trees to predict the outcome as opposed to decision tree regression (single tree). With the help of this algorithm, a decision tree is constructed using k randomly chosen data points from the provided dataset. The worth of any new data point is then predicted using a number of decision trees. A random forest algorithm will forecast multiple output values because there are numerous decision trees. To determine the final result for a new data point, you must discover the average of all the predicted values. This occurs as a result of the numerous decision trees that must be mapped using this method. more processing capacity. Trees run parallel; it's a bagging method, not a boosting technique. i.e., there is no interaction between these trees as you construct trees.

Classification Algorithms:

An algorithm called classification discovers functions to categorise the dataset into groups based on different criteria. On the basis of what it learns from the training dataset, a computer programme divides the data into different groups.

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- KNN - K Nearest Neighbour

Logistic Regression:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Under the category of supervised learning is logistic regression. Using a predetermined collection of independent variables, it is used to predict the categorical dependent variable. In a categorical dependent variable, the outcome is predicted by logistic regression. As a result, the result must be a discrete or categorical number. Rather than providing the precise values of 0 and 1, it provides the probabilistic values that fall between 0 and 1In logistic regression, we fit a "S" shaped logistic function, which forecasts two maximum values, rather than a regression line. (0 or 1). By using various kinds of data to categorise the observations, logistic regression can be used to quickly determine the most efficient factors that were used for classification.

Classification Algorithm (Decision Tree):

A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each child node for the classification result. The Decision Node and Leaf Node are the two elements in a decision tree. While Leaf nodes are the results of decisions and do not have any additional branches, Decision nodes are used to make decisions and have numerous branches. It is a graphical representation for obtaining all feasible answers to a decision or issue based on predetermined conditions. Since it functions like a tree, it is known as a judgement tree. begins with the base node and grows on additional branches to create a structure resembling a tree. The CART algorithm, which means for Classification and Regression Tree algorithm, is used to construct a tree. It can be applied to classification and regression.

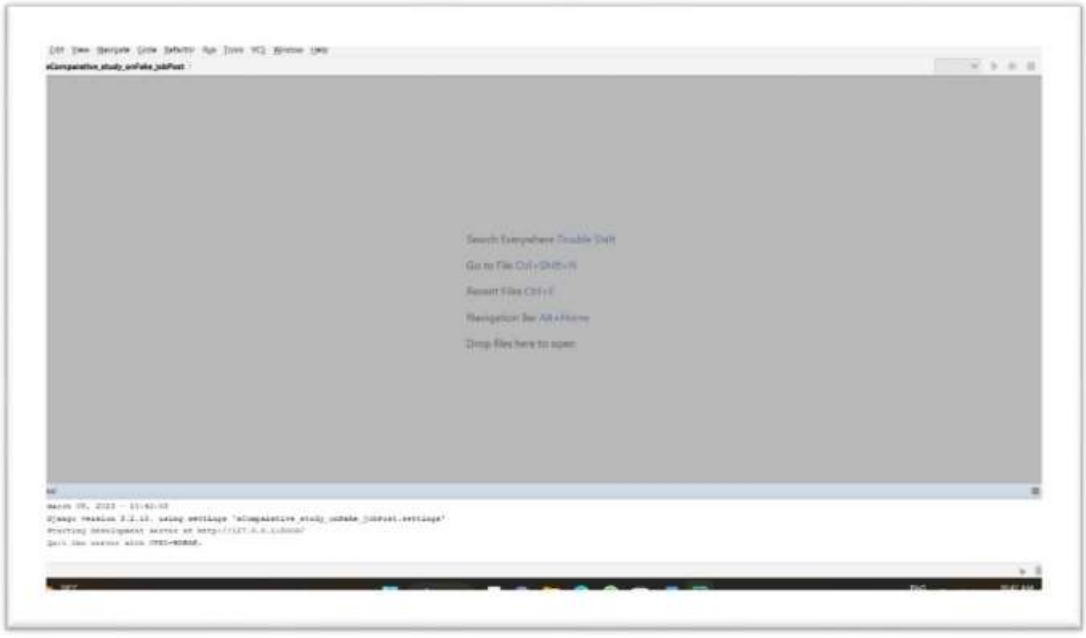
Random Forest Algorithm:

Popular machine learning algorithm Random Forest is a part of the guided learning methodology. In order to increase the dataset's predictive accuracy, a classifier called Random Forest uses multiple decision trees on different segments of the input data. It can be applied to ML issues involving both classification and regression. It is founded on the idea of ensemble learning, which is a method of combining various classifiers to address complex issues and enhance model performance. Even for the large dataset, it operates effectively and predicts the outcome with a high degree of accuracy. When a significant amount of the data is absent, accuracy can still be maintained.**K-Nearest Neighbour (KNN)**

5 RESULTS AND DISCUSSION

SCREENSHOTS

Path setting



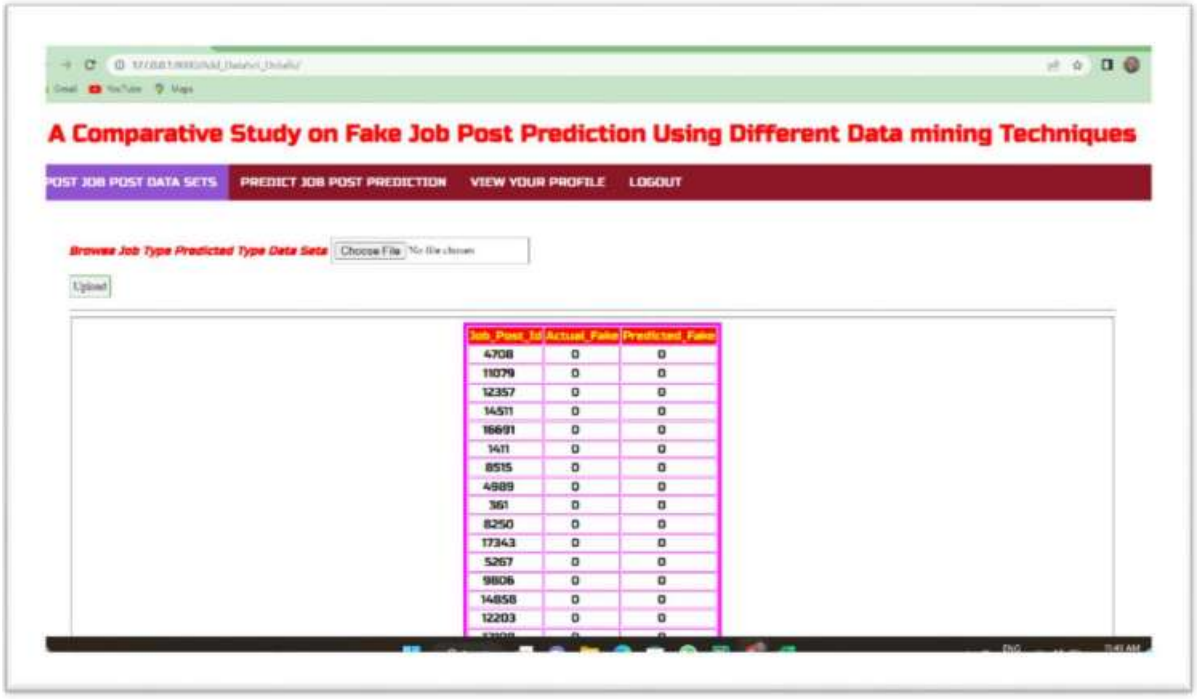
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

REGISTER PAGE

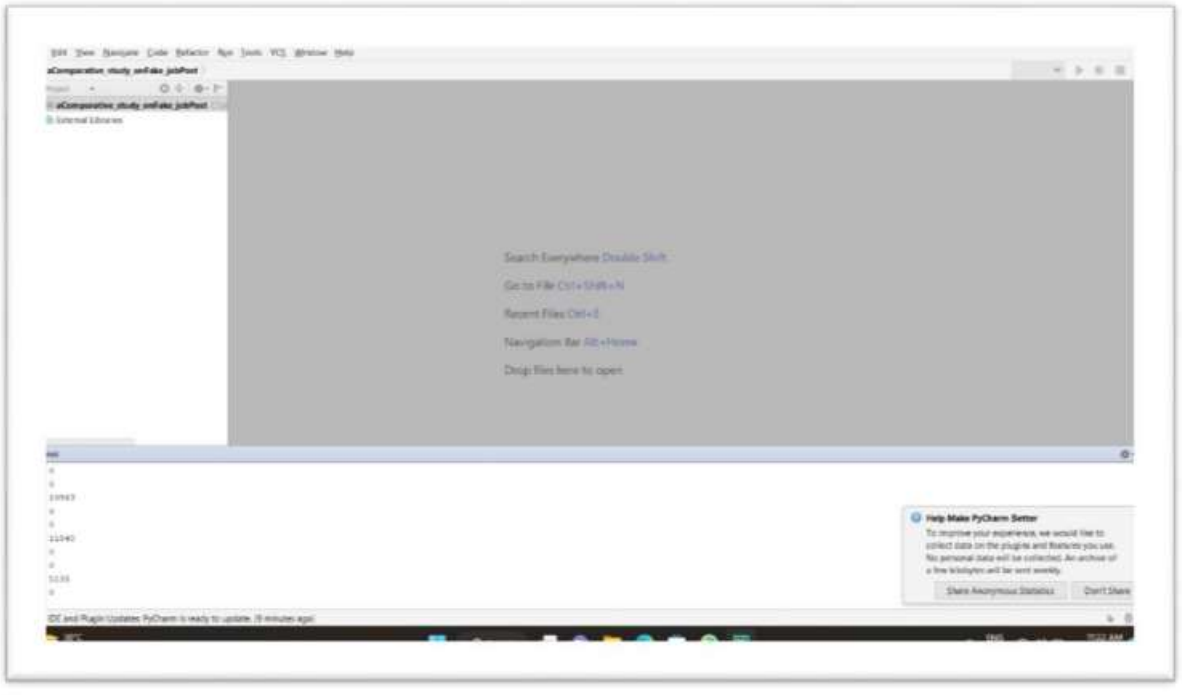
LOGIN PAGE

JOB PREDICTION IN DATA SET

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

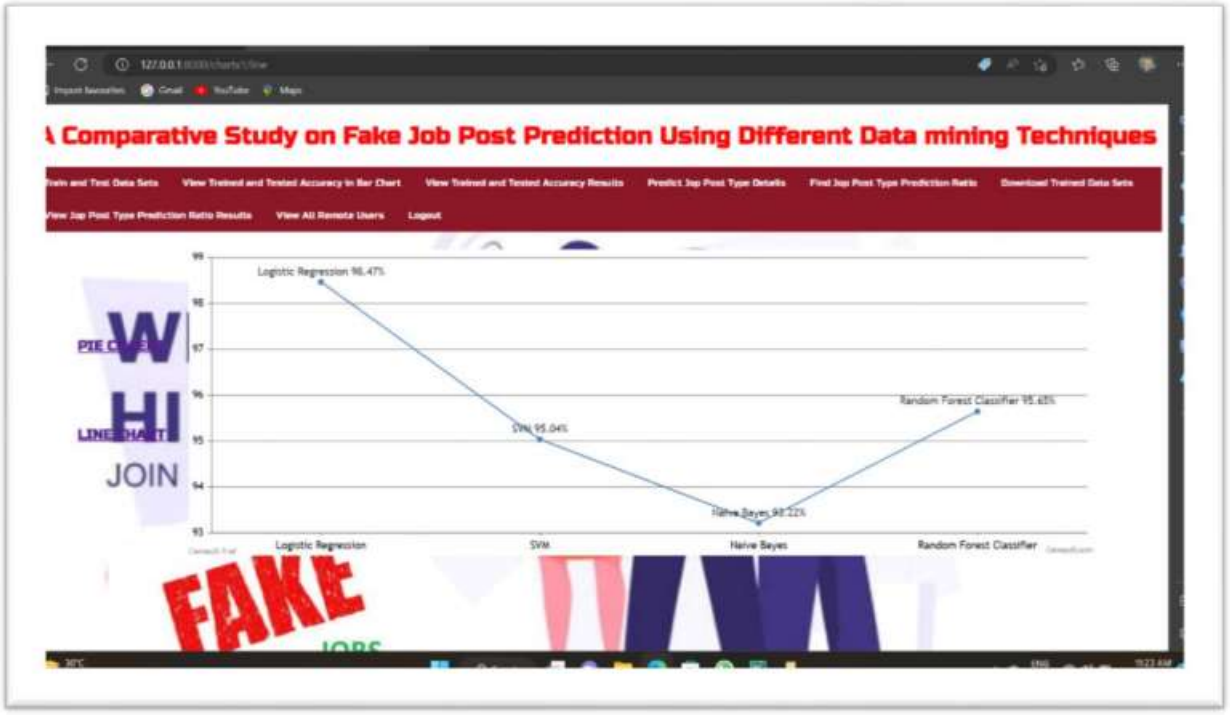


PREDICTION VALUES IN PYCHAM

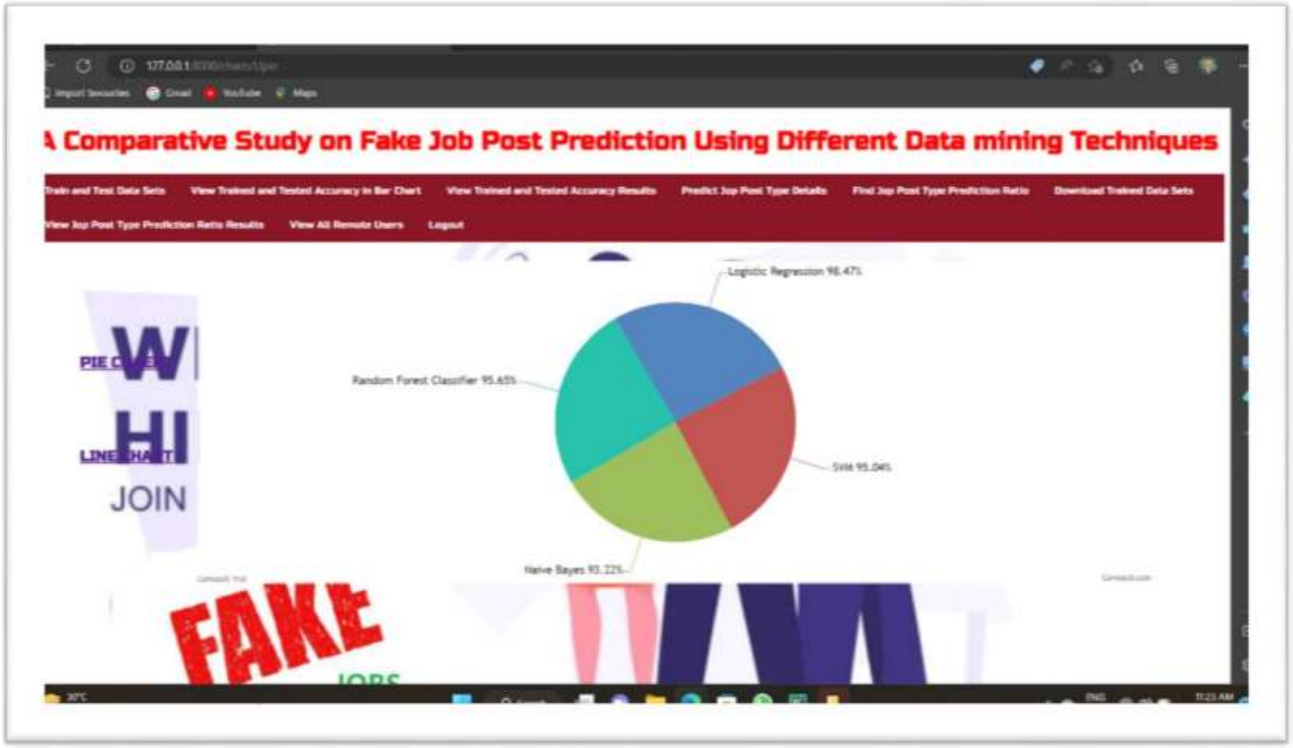


LINE CHART

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

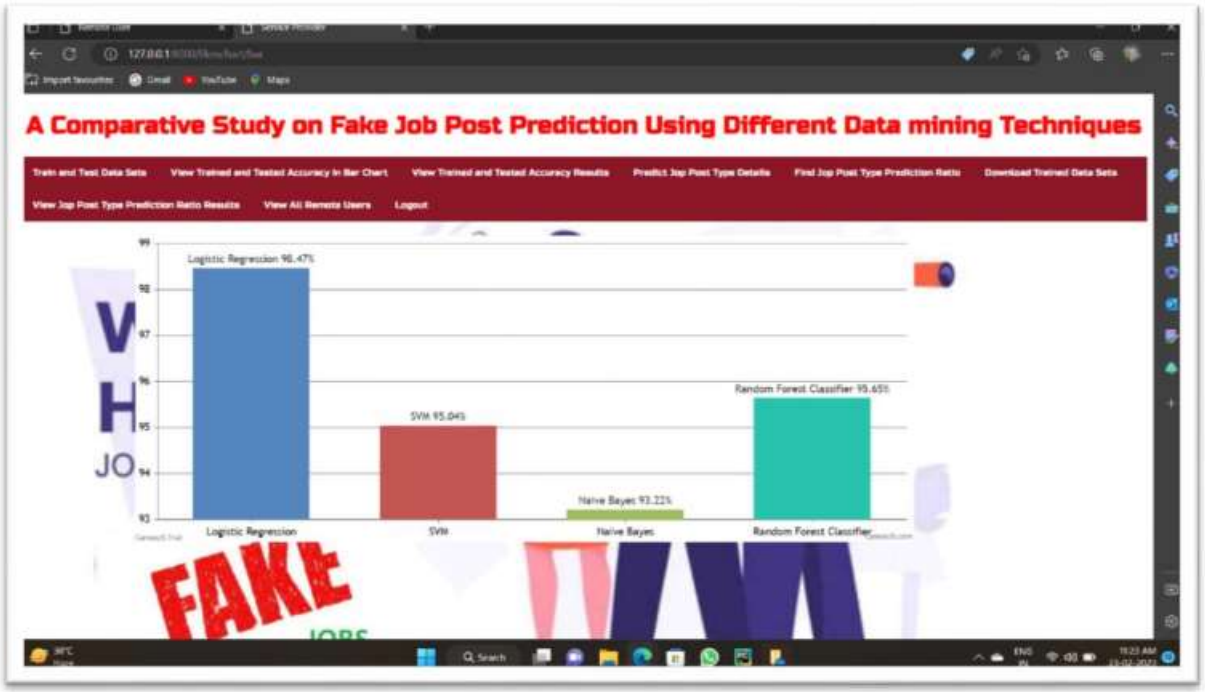


PIE CHART

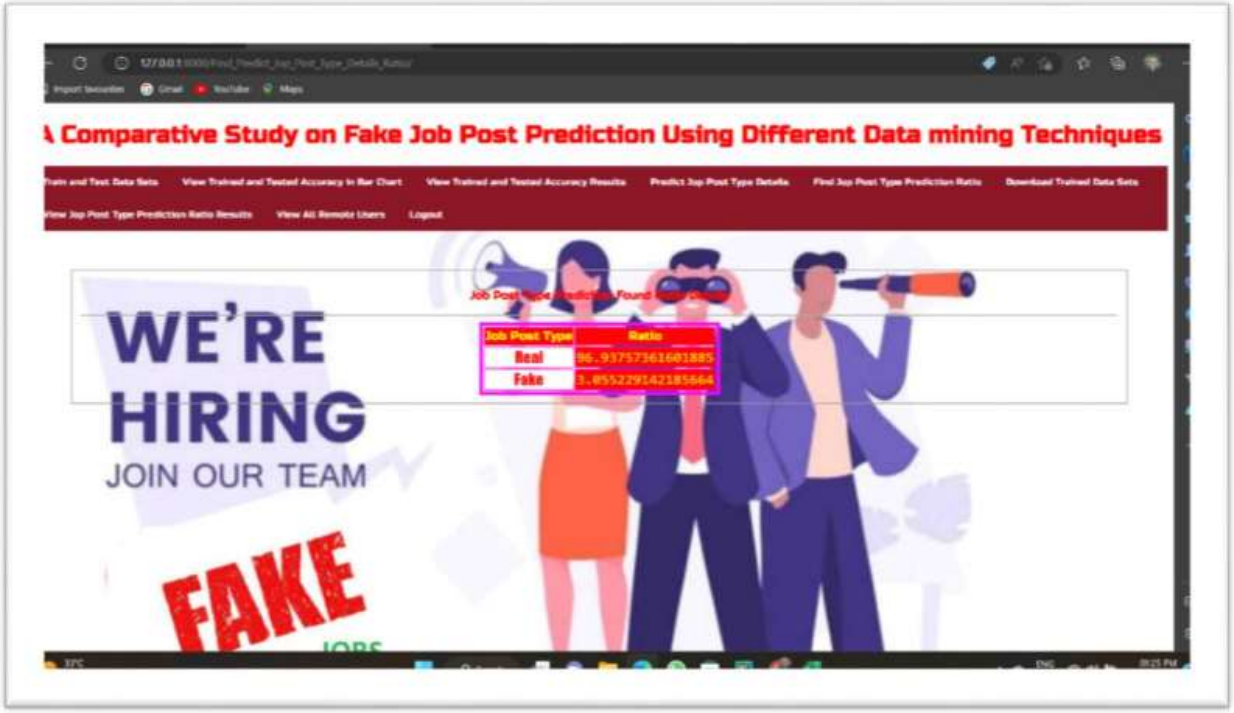


BAR CHART

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

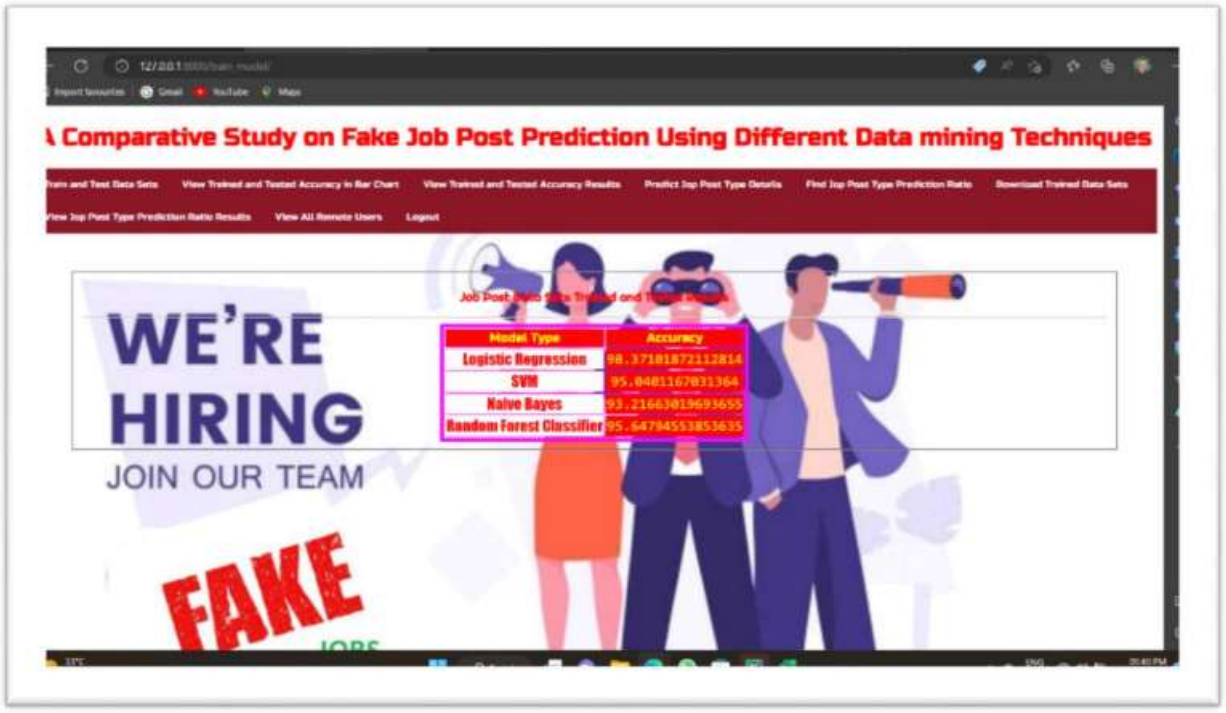


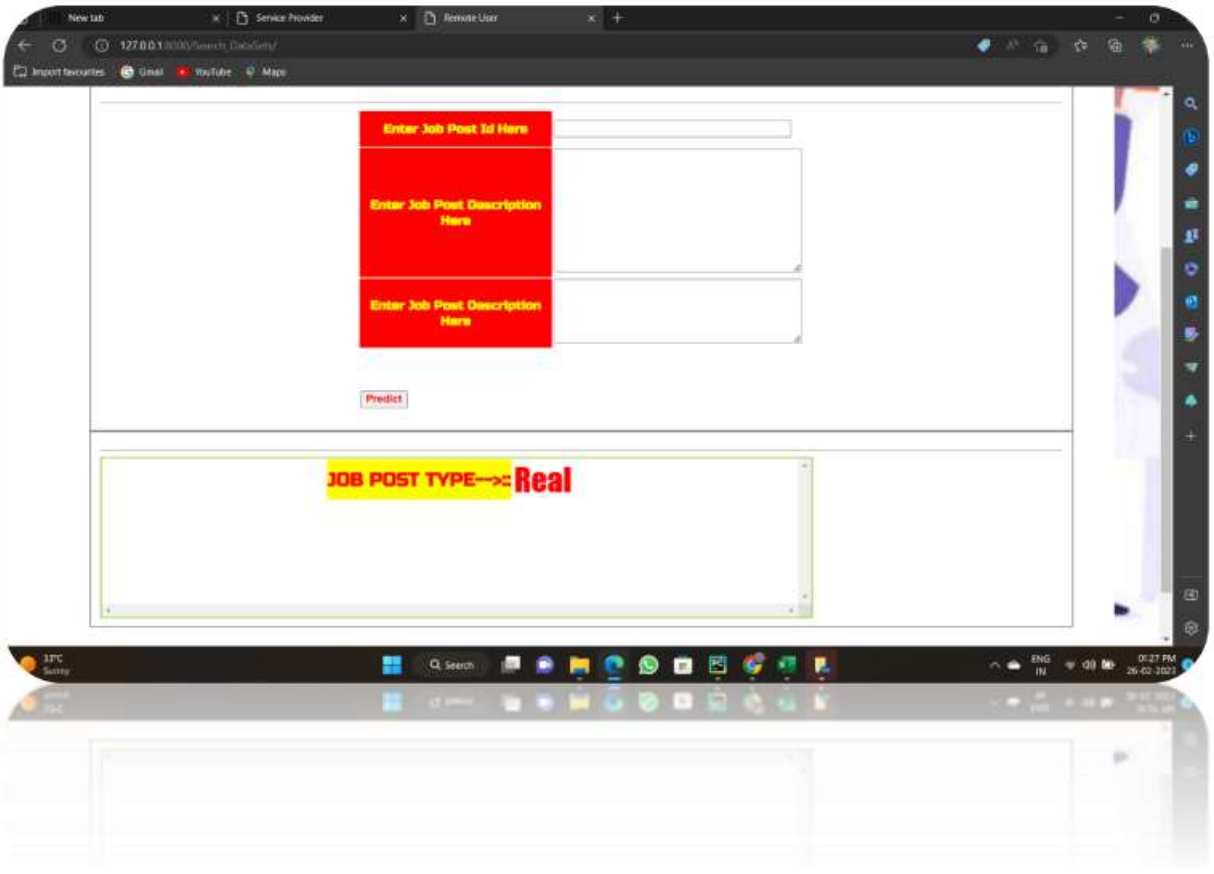
RATIO



RATIO OF ACCURACY

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023





7. CONCLUSION AND FUTURE WORK

The detection of job scams has recently become a major problem worldwide. We have examined the effects of employment scams in this paper since they might be a very lucrative topic of study and make it difficult to identify fake job postings. We conducted experiments using the EMSCAD dataset, which includes real-world fictitious job postings. In this study, we tested both deep learning models and machine learning algorithms (SVM, KNN, Nave Bayes, Random Forest, and MLP) (Deep Neural Network). This article presents a comparison study on the assessment of classifiers based on deep learning and conventional machine learning. . Among conventional machine learning methods, Random Forest Classifier has the highest classification accuracy, followed by DNN (fold 9) with 99% accuracy and Deep Neural Network with an average classification accuracy of 97.7%.

8. REFRENCES

[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 with Deep 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," arXiv Prepr. 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," arXiv Prepr. 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.

[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. https://doi.org/10.5121/imsa.2016.8405

[12] Von Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen "Emotion Recognition for Vietnamese Social Media Text", arXiv Prepr. arXiv:1911.09339, 2019.

[13] 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.

[14] Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14-17 December 2014; pp. 899-904.

[15] Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online review communities. In Proceedings of the 21st international conference on World Wide Web, Lyon, France, 16-20 April 2012; ACM: New York, NY, USA, 2012; pp. 201-210.

[16] Nizam ani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature Abundance. Egyptian Informatics Journal, Vol.15, pp.169-174