Secure Hire: Ml-Powered Detection of Fake Job Advertisements

M. K.D.S. Prasadu[1], Ch.K.V.S.Bhanu Sri[2], Ch.Shalini Priya[3], G.Jhansi[4], K.Varshini[5], K.Satya[6]

[1]Assistant Professor, CSE Department, Vsm College Of Engineering, Ramachandrapuram, Dr.B.R.Ambedkar Konaseema Dt, Andhra Pradesh - 533255
[2,3,4,5,6]Students, CSE Department, Vsm College Of Engineering, Ramachandrapuram, Dr.B.R.Ambedkar Konaseema Dt, Andhra Pradesh - 533255

ABSTRACT
The research paper describes an automated solution that uses machine learning to recognize fake job postings on the internet. The best employment scam detection model is selected by comparing the results of various classifiers used to detect fraudulent posts on the internet. It facilitates in identifying fraudulent job postings among a large number of candidates. There are two types of classifiers used to detect fraudulent job posts: individual classifiers and ensemble classifiers. Experiments show that ensemble classifiers are more effective at detecting frauds than individual classifiers.

Keywords: Genuine (or) Fake Job, Online Recruitment, Machine Learning, Natural Language Processing (NLP), Ensemble Classifiers and Supervised Algorithms.

INTRODUCTION
Online Recruitment Frauds (ORF) has recently addressed the critical issue of employment scams . Companies are increasingly posting job openings online for easy and rapid access by job searchers. Scammers may offer job-seekers jobs in exchange for money, which might be considered a scam. Posting fraudulent job adverts might harm a reputable companies credibility. Automated tools can recognize and report false job postings, helping people avoid applying for them.
To detect fake posts, a machine learning approach is applied using various classification algorithms. A classification tool detects fake job postings and alerts the user. To detect job scams, supervised learning algorithms are proposed as a primary classification technique. Classifiers use training data to map input variables to target classes. The paper briefly describes the classifiers used to distinguish between false and genuine job postings. Classifier-based prediction can be divided into two categories: single classifiers and ensemble classifiers.
For this purpose, machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques are considered initially. A classifier maps input variable to target classes by considering training data. Classifiers addressed in the paper for identifying fake job posts from the others are described briefly. These classifiers based prediction may

be broadly categorized into Single Classifier based Prediction and Ensemble Classifiers based Prediction.

A) Single Classifier based Prediction: Classifiers are trained to predict unknown test instances. The following classifiers are used to detect fake job posts.

a) Naive Bayes Classifier:

This supervised classification method uses the Bayes Theory of Conditional Probability. This classifier's judgment is highly effective in practice, even though its probability estimates are incorrect. This classifier produces a very good outcome in the following instance. When features are independent or entirely functionally based. This classifier's performance is determined by the amount of information lost due to the independence assumption, not by feature dependencies.

b) Decision Tree Classifier:-

A Decision Tree (DT) is a classifier that best illustrates the application of a tree-like structure. It learns more about classification. Every target class is represented as a leaf node in a decision tree, and non-leaf nodes are used as decision nodes that represent particular tests. Any of the decision node's branches can determine the results of those tests.

c) Logistic Regression Classifier:-

A statistical model used for binary classification tasks is called logistic regression. By applying the logistic function to a linear combination of the input attributes, it determines the likelihood that a given input is a member of a particular class.

Ensemble Classifier based Prediction:

The ensemble approach makes it easier for multiple machine learning algorithms to work together to increase the system's overall accuracy.

d) Random Forest Classifier:-

Random forest (RF) makes use of the ensemble learning technique, which proves beneficial for problems including classification. This classifier combines a number of tree-like classifiers, each of which votes for the class that best fits the input by applying its weight to different dataset subsamples.

EXISTING SYSTEM

In order to identify fraudulent job postings, the existed system for job advertisements usually depends on manual examination and human judgment. This procedure can be challenging and prone to mistakes, which can result in inefficiencies and put job searchers at risk of being target of frauds. Further, the absence of automated tools increases the possibility that false advertisements for jobs may go unnoticed, which is an ongoing challenge to the recruitment process quality.

In manual evaluation is subjective, different reviewers may recognize fraudulent job posts differently, which can result in inconsistent and inaccurate scam identification. In order to find common qualities, such as job descriptions that don't match or instant employment without requiring an interview, the system use pattern matching techniques.

PROPOSED SYSTEM

The aim of this research is to figure out if a job posting is fraudulent, Finding and removing these false job postings will make it easier for job searchers to focus only on real job postings. Here, data from a Kaggle dataset is used to provide details about a job that could either be genuine or fraud. The schema of the dataset is displayed in Fig 1.

Parameters	Values
Job Title	object
Job Location	object
Department	object
Salary Range	int64
Description	object
Company Profile	object
Requirements	object
Required Education	object
Required Experience	object
Benefits	object
Employment Type	object
Industry	object
Function	object

Fig 1. Schema Structure of the Dataset

In this collection, 17,880 job postings are present. The suggested methods for evaluating the approach's overall performance make use of this dataset. A multistep approach is used to produce a balanced dataset, assisting in a better understanding of the goal as a standard. This dataset undergoes some pre-processing methods before being fitted to a classifier. Pre- processing procedures includes the removal of stop words, unnecessary attributes, irrelevant spaces, and missing values. This provides the dataset ready to be converted to a categorical encoding format so that a feature vector may be obtained. Several classifiers are fitted with these feature vectors. The working paradigm of a classifier for prediction is described in the following diagram Fig 2.

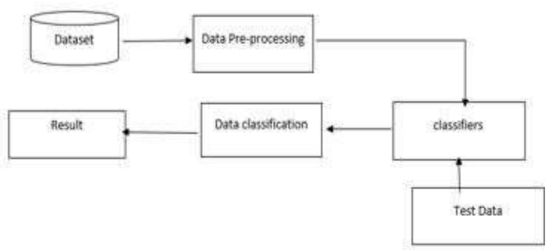

Fig 2. Working of Classifiers

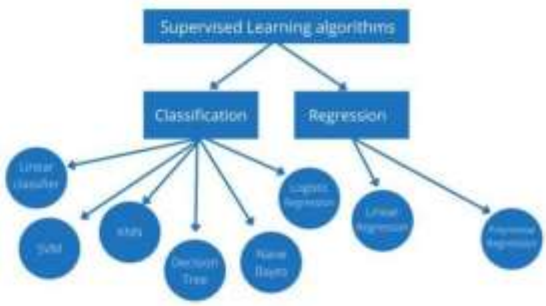

Fig 3. Classification models

A few classifiers, such as Naive Bayes classifier, Decision Tree classifier, Logistic Regression classifier, and Random Tree classifier, are used to identify which job posts are fake, as shown in Fig 3. Note that for classification purposes, the dataset's attribute "fraudulent" is retained as the target class. Initially, 80% of the whole dataset is utilized to train the classifiers, and thereafter, 20% of the dataset is used for prediction. The prediction for each of these classifiers is assessed using performance metrics like accuracy, precision, recall and F1_score. Ultimately, the optimal candidate model is determined by selecting the classifier that performs the best across all metrics. Ultimately, the best candidate model for the classifier is determined by evaluating its performance across all metrics.

Implementation of Classifiers:- Classifiers are trained with suitable parameters in this framework. The default parameters may not be sufficient to maximize the performance of these models. The model can be considered optimal for separating out fraudulent job advertisements from job seekers by adjusting these parameters, which improves the model's dependability. However, 500 estimators are used in the construction of ensemble classifiers such as Random Forest classifier. These classification models are created and then training data are fitted into them. The testing dataset is then utilized to make predictions. Following prediction, the classifiers performance is assessed using the difference between the actual and predicted values.

Performance Evaluation Metrics:- It is crucial to use certain metrics while assessing a model's performance skill in order to validate the evaluation. The following metrics are used for this purpose in order to determine the most suitable and appropriate method of problem-solving. A metric called accuracy measures the proportion of accurate forecasts to all instances that were taken into consideration. However, as the accuracy does not account for incorrectly predicted situations, it could not be a sufficient indicator to assess the performance of the model. There is a serious issue when a fraudulent post is taken seriously. As a result, instances of false positives and false negatives that offset misclassification must be taken into consideration. It is essential to take precision and recall into consideration while evaluating this model.

The ratio of correctly predicted positive results to the total number of positive results by the classifier is known as precision. Recall is calculated by dividing the total number of relevant samples by the number of correct positive outcomes. The harmonic mean of recall and precision is used to compute the F1-Score, also known as the F-measure, which is a metric that takes into account both recall and accuracy. This metric, which determines inter- rate agreement for qualitative elements in a classification problem, is a statistical measure. Another assessing metric that calculates the absolute differences between the test sample observations and predictions is Mean Squared Error (MSE). A model that performs better is shown by a lower mean squared error (MSE) and greater accuracy and F1-score values.

EXPERIMENTAL RESULTS

Over a given dataset that comprises both fake and real posts, all of the above mentioned classifiers are trained and tested for the purpose of detecting fake job postings. Table 1 presents a comparative analysis of the single classifiers in terms of evaluation metrics, while Table 2 presents the outcomes for the classifiers that depend on ensemble approaches. The overall performance of each classifier is shown in Fig 4, 5, and 6 in terms of accuracy, f1-score, and mean squared error, respectively.

TABLE I

PERFORMANCE COMPARISION CHART FOR SINGLE CLASSIFIER BASED PREDICTION

Performan ce Measure Metric	Naive Bayes Classifier	Logistic Regression Classifier	Decision Tree Classifier
Accuracy	72.06%	96.14%	97.2%
F1_Score	0.72	0.96	0.97
MSE	0.52	0.05	0.03

TABLE II
PERFORMANCE COMPARISION CHART FOR ENSEMBLE CLASSIFIER BASED PREDICTION

Performance Measure Metric	ındom Forest Classifier
Accuracy	98.27%
F1-Score	0.97
MSE	0.02

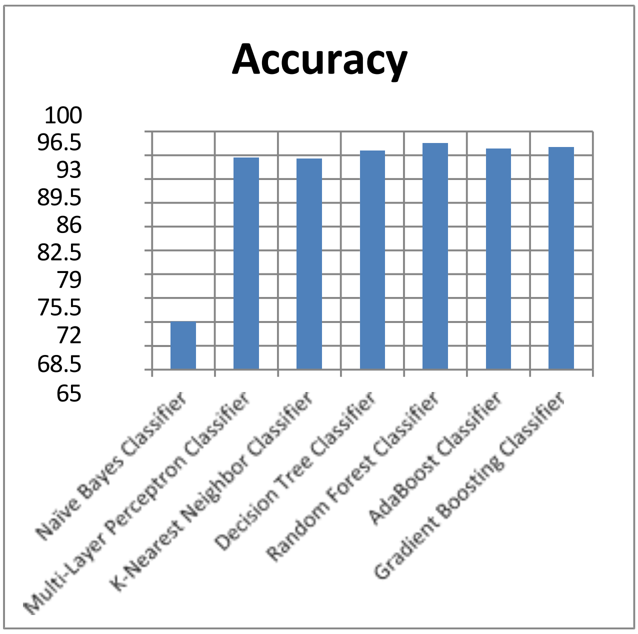

Fig 4. Comparison of Accuracy for all specified supervised machine learning model.

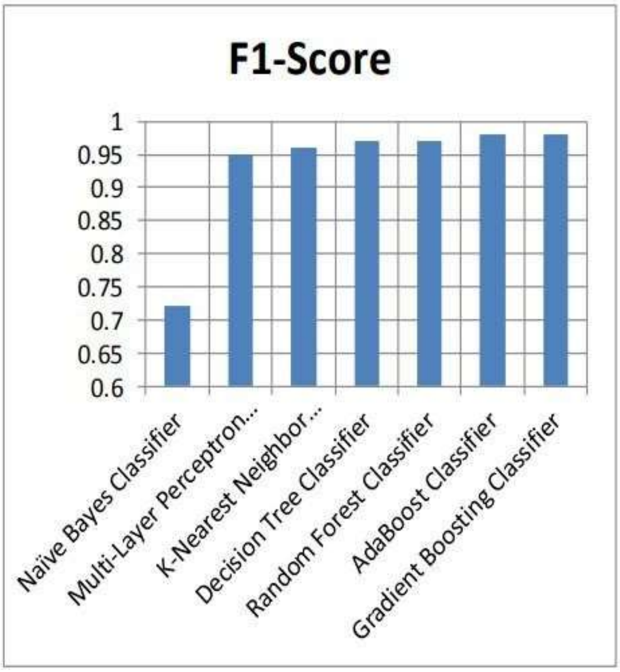

Fig 5. Comparison of F1_Score for all specified supervised machine learning model.

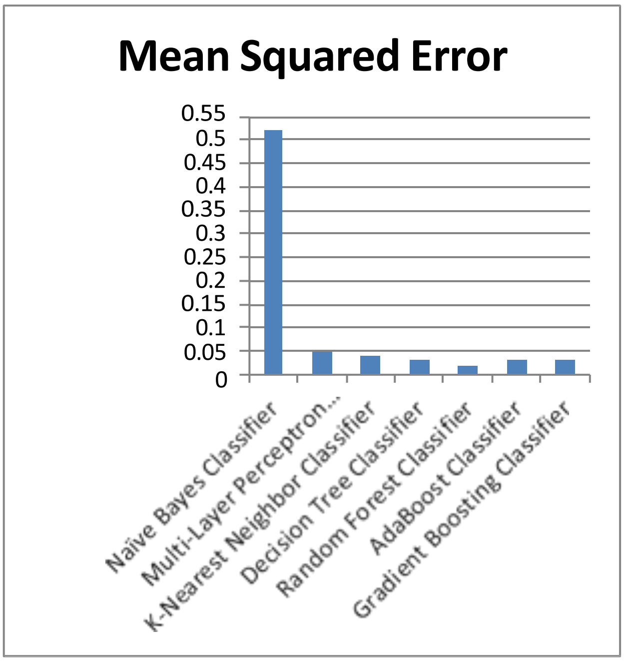

Fig 6. Comparison of MSE for all specified supervised machine learning model.

Table 1 clearly shows that the Decision Tree Classifier exceeds the Naïve Bayes and Logistic Regression classifiers in terms of promising results. As a result, the Decision Tree Classifier simply can be a successful predictor. It is now being determined if using the ensemble approach improves the model's performance or not. Because of this, Random Tree classifiers are used and evaluated in relation to metrics. According to experimental findings, ensemble-based classifiers outperform the other models shown in Table 1 in terms of performance. However, Table 2 shows that because the Random Tree classifier uses many Decision Tree classifiers, it performs significantly better than other similar models. While the Random Forest Classifier is also effective, it is evident that the Decision Tree classifier is the most competitive among its competitors. This classifier has a 98.27% accuracy rate, an F1-score of 0.97, and an MSE of 0.02. Despite having an F1-score that is nearly identical to that of other similar classifiers, this Random Forest classifier displayed significant performance in terms of other measures. As a result, the Random Forest classifier is thought to be the most effective model for this fraudulent job verification system.

CONCLUSION

Identifying employment scams will help job seekers receive only genuine offers from companies. This research proposes different machine learning algorithms as solutions to address job scam detection. To demonstrate the use of several classifiers for job fraud detection, supervised technique is employed. The Random Forest classifier performs better than its peer classification tool, according to experimental results. Compared to the current methods, the accuracy of the proposed methodology was 98.27%, which is significantly better.

REFERENCES

1. Bandar Alghamdi, Fahad Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, 2019, pp. 155-176.
2. Tao Jiang, Jian ping li, Amin ul Haq, Abdus labor, and Amjad al, “A Novel Stacking Approach for Accurate Detection of Fake News”, Vol. 9, 2021, pp. 22626-22639.
3. Karri sai Suresh reddy, karri Lakshmana reddy, “fake job recruitment detection”, JETIR August 2021, Vol. 8, pp. d443-d448.
4. Tulus Suryanto, Robbi Rahim, Ansari Saleh Ahmar, “Employee Recruitment Fraud Prevention with the Implementation of Decision Support System”, Journal of Physics Conference Series, 2018, pp.1-11.
5. C. Jagadeesh, Dr. Pravin R Kshirsagar, G. Sarayu, G.Gouthami, B.Manasa, “Artificial intelligence based Fake Job Recruitment Detection Using Machine Learning Approach”, Journal of Engineering Sciences, Vol. 12, 2021, pp. 0377-9254.
6. Lal, Sangeeta, Rishabh Jiaswal, Neetu Sardana, Ayushi Verma, Amanpreet Kaur, and Rahul Mourya. "ORFDetector: ensemble learning based online recruitment fraud detection." In 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1-5. IEEE, 2019.
7. Samir Bandyopadhyay, Shawni Dutta, “Fake Job Recruitment Detection Using Machine Learning Approach”, International Journal of Engineering Trends and Technology (IJETT),Vol. 68, 2020, pp. 48- 53.
8. George Tsakalidis, Graduate Student Member, IEEE, and Kostas Vergidis, “A Systematic Approach Toward Description and Classification of Cybercrime Incidents”, IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 49, 2019, pp. 1-20.