International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Intelligent Machine learning model to predict fake job posting

P. VEERANNA, Department Of IT, SICET, Hyderabad

MACHA LOKESHWARI, SHESHAGONI RAJASREE, KALIDINDI SAI SHASHANK VARMA, GUJJA BHAVANI

UG Student, Department Of IT, SICET, Hyderabad

Abstract *— To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed in the paper. Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts. Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers.*

Keywords — *Fake Job, Online Recruitment, Machine Learning, Ensemble Approach.*

I. INTRODUCTION

Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF) [1]. In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the job-seekers. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. Fraudulent job advertisements can be posted against a reputed company for violating their credibility. These fraudulent job post detection draws a good attention for obtaining an automated tool for identifying fake jobs and reporting them to people for avoiding application for such jobs.

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 for predicting the unknown test cases. The following classifiers are used while detecting fake job posts-

a) Naive Bayes Classifier-

The Naive Bayes classifier [2] is a supervised classification tool that exploits the concept of Bayes Theorem [3] of Conditional Probability. The decision made by this classifier is quite effective in practice even if its probability estimates are inaccurate. This classifier obtains a very promising result in the following scenario- when the features are independent or features are completely functionally dependent. The accuracy of this classifier is not related to feature dependencies rather than it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy [2].

b) Multi-Layer Perceptron Classifier-

Multi-layer perceptron [4] can be used as supervised classification tool by incorporating optimized training parameters. For a given problem, the number of hidden layers in a multilayer perceptron and the number of nodes in each layer can differ. The decision of choosing the parameters depends on the training data and the network architecture [4].

c) K-nearest Neighbor Classifier-

K-Nearest Neighbour Classifiers [5], often known as lazy learners, identifies objects based on closest proximity of training examples in the feature space. The classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on choosing the appropriate value of k [5].

d) Decision Tree Classifier-

A Decision Tree (DT) [6] is a classifier that exemplifies the use of tree-like structure. It gains knowledge on classification. Each target class is denoted as a leaf node of DT and non-leaf nodes of

DT are used as a decision node that indicates certain test. The outcomes of those tests are identified byeither of the branches of that decision node. Starting from the beginning at the root this tree are goingthrough it until a leaf node is reached. It is the way of obtaining classification result from a decision tree [6].Decision tree learning is an approach that has been applied to spam filtering. This can be useful for forecasting the goal based on some criterion by implementing and training this model [7].

B. Ensemble Approach based Classifiers-

Ensemble approach facilitates several machine learning algorithms to perform together to obtain higher accuracy of the entire system. Random forest (RF) [8] exploits the concept of ensemble learning approach and regression technique applicable for classification based problems. This classifier assimilates several tree-like classifiers which are applied on various sub-samples of the dataset andeach tree casts its vote to the most appropriate class for the input.

Boosting is an efficient technique where several unstable learners are assimilated into a single learner in order to improve accuracy of classification [9]. Boosting technique applies classification algorithm to the reweighted versions of the training data and chooses the weighted majority vote of the sequence of classifiers. AdaBoost [9] is a good example of boosting technique that produces improved output even when the performance of the weak learners is inadequate. Boosting algorithms are quite efficient is solving spam filtration problems. Gradient boosting [10] algorithm is another boosting technique based classifier that exploits the concept of decision tree. It also minimizes the prediction loss.

II. RELATED WORK

According to several studies, Review spam detection, Email Spam detection, Fake news detection have drawn special attention in the domain of Online Fraud Detection.

A. Review Spam Detection-

People often post their reviews online forum regarding the products they purchase. It may guide other purchaser while choosing their products. In this context, spammers can manipulate reviews for gaining profit and hence it is required to develop techniques that detects these spam reviews. This can be implemented by extracting features from the reviews by extracting features using Natural Language Processing (NLP). Next, machine learning techniques are applied on these features. Lexicon based approaches may be one alternative to machine learning techniques that uses dictionary or corpus to eliminate spam reviews[11].

B. Email Spam Detection-

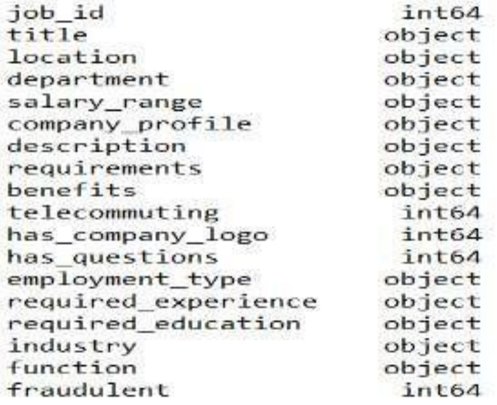
Unwanted bulk mails, belong to the category of spam emails, often arrive to user mailbox. This may lead to unavoidable storage crisis as well as bandwidth consumption. To eradicate this problem, Gmail, Yahoo mail and Outlook service providersincorporate spam filters using Neural Networks. While addressing the problem of email spam detection, content based filtering, case based filtering,heuristic based filtering, memory or instance based filtering, adaptive spam filtering approaches are taken into consideration [7].

C. Fake News Detection-

Fake news in social media characterizes malicious user accounts, echo chamber effects. Thefundamental study of fake news detection relies on three perspectives- how fake news is written, how fake news spreads, how a user is related to fake news.Features related to news content and social contextare extracted and a machine learning models are imposed to recognize fake news [12].

III. PROPOSED METHODOLOGY

The target of this study is to detect whether a job post is fraudulent or not. Identifying and eliminating these fake job advertisements will help the job- seekers to concentrate on legitimate job posts only. In this context, a dataset from Kaggle [13] is employed that provides information regarding a job that may or may not be suspicious. The dataset has the schema as shown in Fig. 1.



```
job_id                    int64
title                    object
location                 object
department               object
salary_range             object
company_profile          object
description              object
requirements             object
benefits                 object
telecommuting             int64
has_company_logo          int64
has_questions             int64
employment_type          object
required_experience      object
required_education       object
industry                 object
function                 object
fraudulent                int64
```

Fig. 1. Schema structure of the dataset

This dataset contains 17,880 number of job posts. This dataset is used in the proposed methods for testing the overall performance of the approach. For better understanding of the target as a baseline, a multistep procedure is followed for obtaining a balanced dataset. Before fitting this data to any classifier, some pre-processing techniques are applied to this dataset. Pre-processing techniques include missing values removal, stop-words elimination,irrelevant attribute elimination and extra space removal. This prepares the dataset to be transformed into categorical encoding in order to obtain a feature vector. This feature vectors are fitted to several classifiers. The following diagram Fig. 2 depicts a description of the working paradigm of a classifier for prediction.

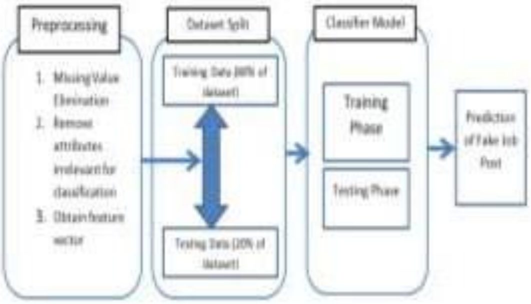

Preprocessing
1. Missing Value Elimination
2. Remove attributes irrelevant for classification
3. Obtain feature vector
Dataset Split
Training Data (80% of dataset)
Testing Data (20% of dataset)
Classifier Model
Training Phase
Testing Phase
Prediction of Fake Job Post

Fig. 2.Detailed description for working of Classifiers

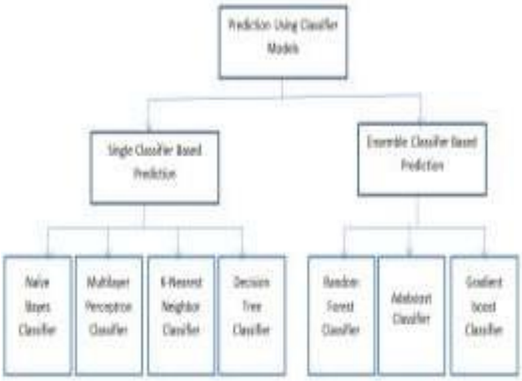

Prediction Using Classifier Models
Single Classifier Based Prediction
Ensemble Classifier Based Prediction
Naive Bayes Classifier
Multilayer Perceptron Classifier
K-Nearest Neighbor Classifier
Decision Tree Classifier
Random Forest Classifier
Adaboost Classifier
Gradient boost Classifier

Fig. 3. Classification models used in this framework

As depicted in Fig. 3, a couple of classifiers are employed such as Naive Bayes Classifier, Decision Tree Classifier, Multi-Layer Perceptron Classifier, K- nearest Neighbor Classifier, AdaBoost Classifier, Gradient Boost Classifier and Random Tree Classifier for classifying job post as fake. It is to be noted that the attribute ‗fraudulent' of the dataset is kept as target class for classification purpose. At first,the classifiers are trained using the 80% of the entire dataset and later 20% of the entire dataset is used for the prediction purpose. The performance measure metrics such as Accuracy, F-measure, and Cohen- Kappa score are used for evaluating the prediction for each of these classifiers. Finally, the classifier that has the best performance with respect to all the metrics is chosen as the best candidate model.

A. Implementation of Classifiers

In this framework classifiers are trained using appropriate parameters. For maximizing the performance of these models, default parameters may not be sufficient enough. Adjustment of these parameters enhances the reliability of this model which may be regarded as the optimised one for identifying as well as isolating the fake job postsfrom the job seekers.

This framework utilised MLP classifier as a collection of 5 hidden layers of size 128, 64, 32, 16 and 8 respectively. The K-NN classifier gives a promising result for the value k=5 considering all the evaluating metric. On the other hand, ensemble classifiers, such as, Random Forest, AdaBoost and Gradient Boost classifiers are built based on 500 numbers of estimators on which the boosting is terminated. After constructing these classification models, training data are fitted into it. Later the testing dataset are used for prediction purpose. After the prediction is done, performance of the classifiers are evaluated based on the predicted value and the actual value.

B. Performance Evaluation Metrics

While evaluating performance skill of a model, it is necessary to employ some metrics to justify the evaluation. For this purpose, following metrics are taken into consideration in order to identify the best relevant problem-solving approach. Accuracy [14] is a metric that identifies the ratio of true predictions over the total number of instances considered.However, the accuracy may not be enough metric for evaluating model‘s performance since it does not consider wrong predicted cases. If a fake post is treated as a true one, it creates a significant problem. Hence, it is necessary to consider false positive and false negative cases that compensate to misclassification. For measuring this compensation, precision and recall is quite necessary to be considered [7].

Precision [14] identifies the ratio of correct positive results over the number of positive results predicted by the classifier. Recall [14] denotes the number of correct positive results divided by the number of all relevant samples. F1-Score or F- measure [14] is a parameter that is concerned for both recall and precision and it is calculated as the harmonic mean of precision and recall [14]. Apart from all these measure, Cohen-Kappa Score [15] is also considered to be as an evaluating metric in this paper. This metric is a statistical measure that finds out inter-rate agreement for qualitative items for classification problem. Mean Squared Error (MSE) [14] is another evaluating metric that measures absolute differences between the prediction and actual observation of the test samples. Lower value of MSE and higher values of accuracy, F1-Score, and Cohen-kappa score signifies a better performing model.

IV. EXPERIMENTAL RESULTS

All the above mentioned classifiers are trained and tested for detecting fake job posts over a given dataset that contains both fake and legitimate posts. The following Table 1 shows the comparative study of the classifiers with respect to evaluating metrics and Table 2 provides results for the classifiers that are based on ensemble techniques. Fig. 4 to Fig. 7 depict overall performance of all the classifiers in terms of accuracy, f1-score, Cohen-kappa score, MSE respectively.

TABLE I
PERFORMANCE COMPARISON CHART FOR SINGLE CLASSIFIER BASED PREDICTION

Performance Measure Metric	Naïve Bayes Classifier	Multi-Layer Perceptron Classifier	K-Nearest Neighbor Classifier	Decision Tree Classifier
Accuracy	72.06%	96.14%	95.95%	97.2%
F1-Score	0.72	0.96	0.96	0.97
Cohen-Kappa Score	0.12	0.3	0.38	0.67
MSE	0.52	0.05	0.04	0.03

TABLE II
PERFORMANCE COMPARISON CHART FOR ENSEMBLE CLASSIFIER BASED PREDICTION

Performance Measure Metric	Random Forest Classifier	AdaBoost Classifier	Gradient Boosting Classifier
Accuracy	98.27%	97.46%	97.65%
F1-Score	0.97	0.98	0.98
Cohen-Kappa Score	0.74	0.63	0.65
MSE	0.02	0.03	0.03

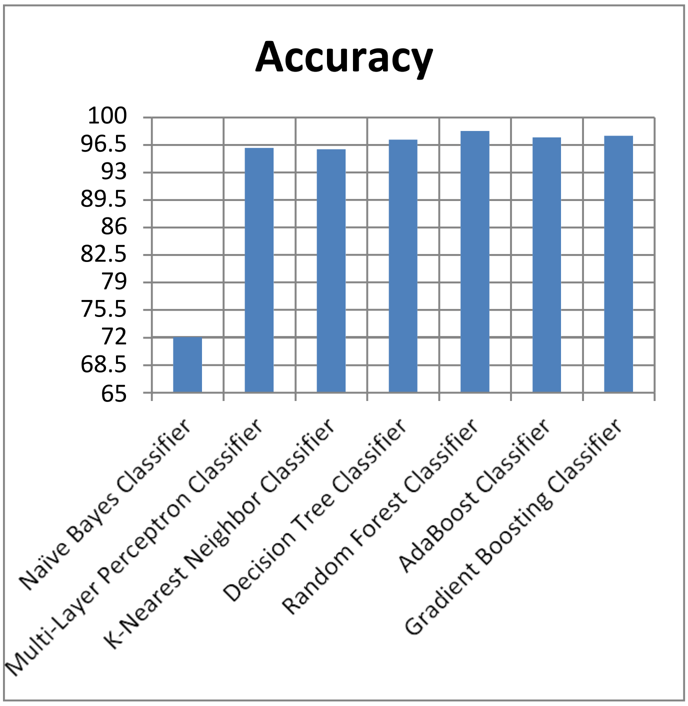


Fig. 4. Comparison of Accuracy for all specified supervised machine learning model

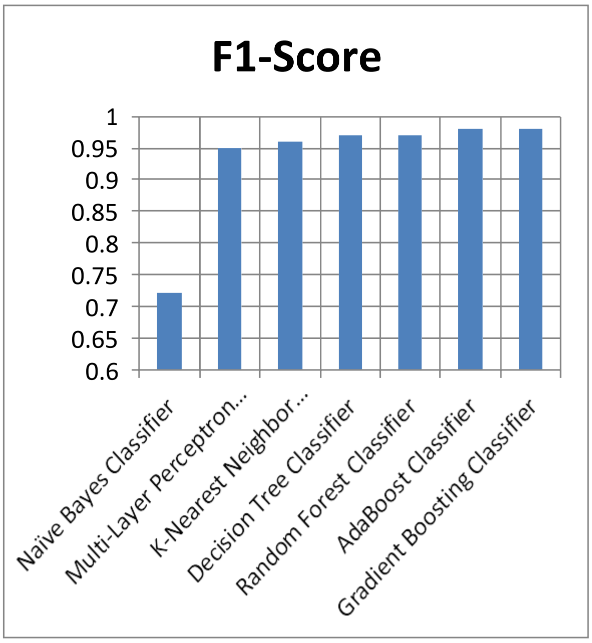


Fig. 5. Comparison of F1-Score for all specified supervised machine learning model

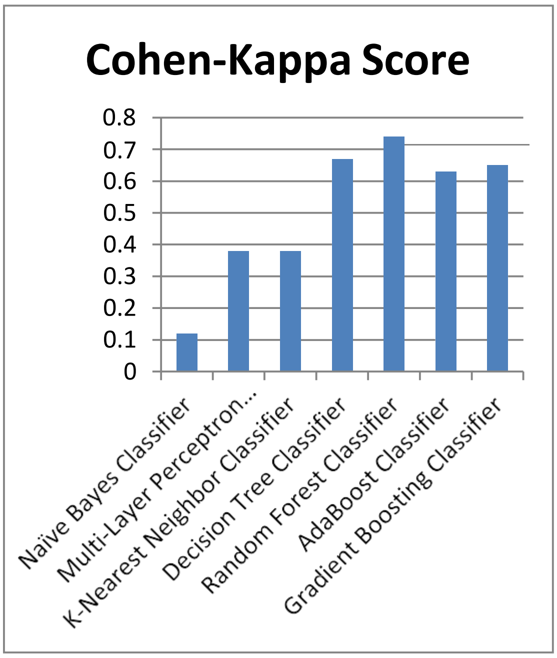


Fig. 6. Comparison of Cohen-Kappa Score for all specified supervised machine learning model

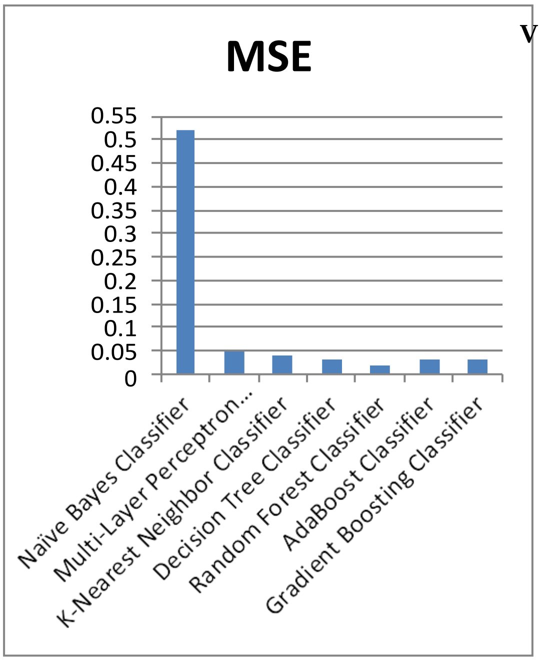


Fig. 7. Comparison of MSE for all specified supervised machine learning model

From Table 1, it is quite clear that Decision Tree Classifier gives promising result over Naïve Bayes Classifier, Multi-Layer Perceptron Classifier, K-Nearest Neighbor Classifier. Hence, Decision Tree Classifier can be fruitful predictor as a single classifier. Now, it is checked whether the use of ensemble approach enhances the performance of the model or not. For that reason, Random Tree classifiers, AdaBoost classifiers and Gradient Boost Classifiers are implemented and compared with respect to the metrics. Experimental results have shown that ensemble based classifiers provide an improved result over the other models specified in Table 1. However, Table 2 indicates that Random Tree classifier outperforms well over its peers because it incorporates multiple Decision Tree classifiers. As it is seen that Decision Tree classifier is the most competing one over its peers, Random Forest Classifier also works well. This classifier has achieved accuracy of 98.27%, Cohen-kappa score as 0.74, F1-score 0.97, MSE 0.02. Though this Random Forest classifier has obtained F1-score which is almost similar to other competitors, but this classifier has shown significant performance with respect to other metrics. Hence Random Forest classifier can be regarded as the best model for this fake job detection scheme.

V. CONCLUSIONS

Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that Random Forest classifier outperforms over its peer classification tool. The proposed approach achieved accuracy 98.27% which is much higher than the existing methods.

REFERENCES

[1] B. Alghamdi and F. Alharby, –*An Intelligent Model for Online Recruitment Fraud Detection,"* J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.

[2] I. Rish, –*An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier*,‖ no. January 2001, pp. 41–46, 2014.

[3] D. E. Walters, –*Bayes's Theorem and the Analysis of Binomial Random Variables*,‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[4] F. Murtagh, –*Multilayer perceptrons for classification and regression*,‖ Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[5] P. Cunningham and S. J. Delany, –*K -Nearest Neighbour Classifiers*,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[6] H. Sharma and S. Kumar, –*A Survey on Decision Tree Algorithms of Classification in Data Mining*,‖ Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[7] E. G. Dada, J. S. Bassi, H. Chiroma, *S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open researchproblems*,‖ Heliyon, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[8] L. Breiman, –*ST4_Method_Random_Forest*,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[9] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli,–*Bagging classifiers for fighting poisoning attacks in adversarial classification tasks,"* Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.

[10] A. Natekin and A. Knoll, –*Gradient boosting machines, a tutorial*,‖ Front. Neurorobot., vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

11N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, –*Spam review detection techniques: A systematic literature review*,‖ Appl. Sci., vol. 9, no. 5, pp. 1– 26, 2019, doi: 10.3390/app9050987.

[11] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, –*Fake News Detection on Social Media*,‖ ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[12] Shivam Bansal (2020, February). [Real or Fake] Fake JobPosting Prediction,Version 1.Retrieved March 29,2020 from https://www.kaggle.com/shivamb/real-or-fake-fake- jobposting-prediction

[13] H. M and S. M.N, –*A Review on Evaluation Metrics for Data Classification Evaluations*,‖ Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[14] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, –*Cohen's kappa coefficient as a performance measure for feature selection,"* 2010 IEEE World Congr. Comput. Intell. WCCI 2010, no. May 2016, 2010, doi: 10.1109/FUZZY.2010.5584447.