iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41641

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

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- Available at www.ijraset.com

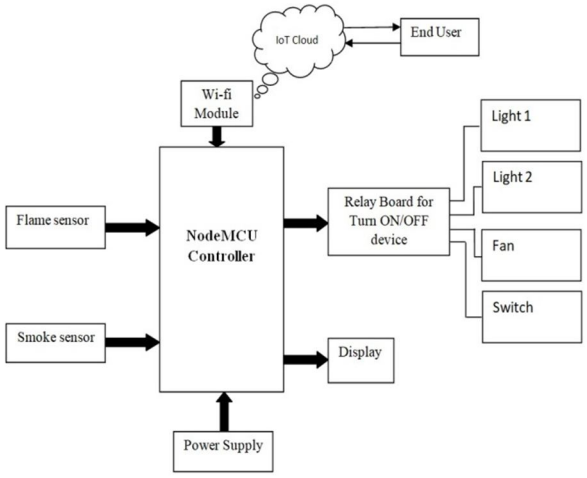
Fake Job Detection Using Machine Learning

Priya Khandagale[1], Akshata Utekar[2], Anushka Dhonde[3], Prof. S. S. Karve[4]
[1, 2, 3]*Datta Meghe College of Engineering Airoli Navi Mumbai, Maharashtra, 400708*
[4]*Guide*

Abstract: ***The research proposes an automated solution based on machine learning-based classification approaches to prevent fraudulent job postings on the internet. Many organizations these days like to list their job openings online so that job seekers may access them quickly and simply. However, this could be a form of scam perpetrated by con artists who offer job seekers work in exchange for money. Many people are duped by this fraud and lose a lot of money as a result. We can determine which job postings are fraudulent and which are not by conducting an exploratory data analysis on the data and using the insights gained. In order to detect bogus posts, a machine learning approach is used, which employs numerous categorization algorithms. The system would train the model to classify jobs as authentic or false based on previous data of bogus and legitimate job postings. To start, supervised learning algorithms as classification techniques can be considered to handle the challenge of recognizing scammers on job postings. It will employ two or more machine learning algorithms, selecting the one that yields the highest accuracy score in the prediction of whether a job advertising headline is genuine or not.***
Keywords: Fake Job, Online Recruitment, Machine Learning, Ensemble Approach.

I. INTRODUCTION

For many people, economic hardship and the impact of the coronavirus have drastically reduced work availability and resulted in job loss. Scammers would love to take advantage of a situation like this. Many individuals are falling prey to these con artists who are preying on people's desperation as a result of an extraordinary event. The majority of fraudsters do this to obtain personal information from the person they are attempting to defraud. Addresses, bank account numbers, and social security numbers are examples of personal information. Scammers provide customers with a fantastic job offer and then demand money in exchange. Alternatively, they may need a financial investment from the job seeker in exchange for the promise of a job. Because of unemployment, there are a lot of job scams these days.



A recruiter can find a qualified candidate through a variety of websites. Fake recruiters will sometimes post a job on a job platform for the sole purpose of making money. Many job boards suffer from this issue. People later go to a new job portal in quest of legitimate employment, but phoney recruiters also migrate to this portal. As a result, it's critical to distinguish between legitimate and fictitious employment opportunities. Employment fraud is one of the most severe concerns that has been addressed in the arena of Online Recruitment Frauds in recent years (ORF). Many organizations these days like to list their job openings online so that job seekers may find them quickly and simply. This could, however, be one form of fraud perpetrated by the con artist. However, this could be a form of scam perpetrated by con artists who offer job seekers work in exchange for money.

This is a dangerous problem that can be solved using machine learning and natural language processing approaches (NLP). In order to detect bogus posts, a machine learning approach is used, which employs numerous categorization algorithms. In this scenario, a classification technique distinguishes bogus job postings from a wider pool of job postings and notifies the user. To start, supervised learning algorithms as classification techniques are being studied to address the challenge of recognizing scammers on job postings. A classifier uses training data to map input variables to target classes. The paper's classifiers for distinguishing phoney job postings from the others are briefly presented. These classifier-based predictions can be divided into two categories: single-classifier predictions and ensemble-classifier predictions.

II. LITERATURE SURVEY

Online recruiting fraud detection is a relatively new sector in which little research has been done. There are some indirect methods to solve Online recruitment fraud to a limited extent, such as Email Spam filtering, which prevents sending advertising-related emails to users, anti-phishing techniques to detect fake websites, and countermeasures against opinion fraud to detect the posting of deceptive and misleading fake reviews. Review spam detection, email spam detection, and fake news identification have all received a lot of attention in the realm of online fraud detection, according to many studies.

A. Review Spam Detection

People frequently share their opinions on the things they buy on online forums. It might be useful to other buyers while they're deciding what to buy. In this setting, spammers can modify reviews for financial advantage, necessitating the development of algorithms to detect spam reviews. This can be done by extracting features from the reviews and using Natural Language Processing to do so (NLP). These features are subjected to machine learning algorithms.

B. Email Spam Detection

Unwanted bulk messages, sometimes known as spam emails, frequently occur in user inboxes. This could result in an inevitable storage shortage as well as increased bandwidth usage. Spam filters based on Neural Networks are used by Gmail, Yahoo Mail, and Outlook to combat this problem. Content-based filtering, case-based filtering, heuristic-based filtering, memory or instance-based filtering, and adaptive spam filtering approaches are all taken into account when tackling the problem of email spam detection.

C. Fake News Detection

In social media, fake news is defined by malicious user accounts and echo chamber effects. Fake news identification is based on three perspectives: how fake news is written, how fake news spreads, and how a user is connected to fake news. To identify fake news, features linked to news content and social context are retrieved, and machine learning algorithms are applied.To the best of the knowledge, Vidros. is the only one who has proposed a strategy for detecting employment fraud. However, they only used a balanced dataset, and the performance of prediction algorithms on an imbalanced dataset has yet to be determined. As a result, evaluating prediction models on an unbalanced dataset is critical. The suggested ORF Detector is an ensemble-based model for detecting online fraud. They applied average vote, majority vote, and maximum vote to three baseline classifiers: J48, Logistic Regression, and Random Forest. However, the fundamental disadvantage of this strategy is that it only works on balanced datasets and produces lower accuracy.

III. SINGLE CLASSIFIER BASED PREDICTION (MODELS IMPLEMENTED)

Unknown test cases are predicted using classifiers that have been learned. When detecting fraudulent job postings, the following classifiers are used-

A. Naive Bayes

The number of parameters required for Naive Bayes classifiers is linear in the number of variables in a learning issue, making them extremely scalable. Instead of expensive iterative approximation, which is used for many other types of classifiers, maximum-likelihood training can be done simply evaluating a closed-form expression in linear time. Naive Bayes is a straightforward method for building classifiers, which are models that give class labels to problem cases represented as vectors of feature values, with the class labels selected from a limited set. The amount of information loss of the class due to the independence assumption is needed to estimate the accuracy of this classifier, not feature dependencies.

B. Support Vector Machine

Support Vector Machines (SVMs) are supervised learning models that can be used to solve tasks like classification and regression. They can solve both linear and nonlinear problems and are useful in a variety of situations. The concept behind Support Vector Machines is straightforward: In a classification problem, for example, the method draws a line between the classes. The line's purpose is to maximise the distance between points on either side of the so-called decision line. After the separation, the model may readily guess the target classes (labels) for new cases, which is a benefit of this procedure.

C. Logistic Regression

It's a categorical response variable that's employed in a classification process. E.g. When the number of hours spent studying is supplied as a feature in predicting whether a student passes or fails an exam, the response variable has two values: pass and fail. Binomial Logistic Regression is a form of issue in which the response variable has two values: 0 and 1, or pass and fail, or true and false. When the response variable can have three or more potential values, Multinomial Logistic Regression is used.

IV. ENSEMBLE APPROACH BASED CLASSIFIERS (RANDOM FOREST)

The ensemble approach allows numerous machine learning algorithms to work together to improve overall system accuracy. Random forest makes use of an ensemble learning approach and a regression technique that can be used to solve classification difficulties.

Random Forest is a classifier that combines a number of decision trees on different subsets of a dataset and averages the results to increase the dataset's predicted accuracy. A random forest is a meta estimator that fits many decision tree classifiers together. Averaging is used to increase forecast accuracy and control over-fitting on various sub-samples of the dataset. If the max samples argument is true, the sub-sample size is restricted; otherwise, the entire dataset is used to create each tree.

This classifier combines many tree-like classifiers, each of which is applied to different sub-samples of the dataset and votes for the most acceptable class for the input.

V. PROPOSED METHODOLOGY

This system's main purpose is to identify whether a job posting is genuine or not. Job seekers will be able to focus entirely on legitimate job openings if fake job postings are identified and deleted. In this system, we plan to use a Kaggle dataset that contains information on the job, including attributes such as job id, title, location, and department. Then there's data preprocessing, which involves removing things like trivial spaces, null entries, stopwords, and so on. The data is provided to the classifier for predictions after it has been preprocessed and cleaned to make it prediction ready.

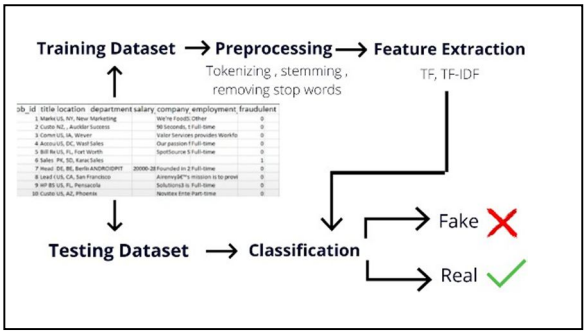


Fig. 1. Fake Job Detection Framework

A. Dataset Details

This kaggle dataset contains 17,880 job posting data entries. We must first preprocess this data in order to prepare it for prediction before fitting it into any of the machine learning models or classifiers. Some pre-processing techniques are used on this dataset before it is fitted to any classifier. Missing values removal, stop-words removal, irrelevant attribute removal, and unnecessary space removal are some of the pre-processing strategies. This prepares the dataset for categorical encoding, which will be used to generate a feature vector.

String	
Title	The title of the job post
Location	Geographical location of the job post
Department	Corporate department (e.g., sales)
Salary range	E.g., $50,000-$60,000
HTML fragment	
Company profile	A short company description
Description	Details of the job ad (post)
Requirements	Required knowledge to apply
Benefits	Employer offered benefits
Binary (true (1), false (0))	
Telecommunicating	True for telecommunicating positions
Company logo	True if there is a company logo
Questions	True if there are questions for applicants
Fraudulent	Classification attribute
Nominal	
Employment type	Full-time, part-time, etc.
Required experience	Entry level intern, etc.
Required education	Bachelor, master, etc.
Industry	IT, health care, etc.
Function	Research, Engineering, etc.

Fig. 2. Detailed description of Data

B. Data Preprocessing

Preprocessing data is the process of transforming raw data into a clean data set. Before running the algorithm, the dataset is preprocessed to check for missing values, noisy data, and other irregularities. It also removes noise and uninformative characters and words from the text, as well as stop-words, extraneous attributes, and extra space. Because of the data set's nature, it needed to be pre-processed before being fed into the classifier. Because the data is textual, we must transform it to a numerical representation before we can make any predictions. Here NLP is used. NLP (Natural Language Processing) **Natural language processing** (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. A pipeline is created to make the textual data machine understandable.

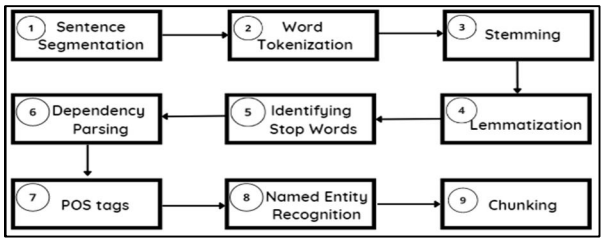


Fig. 3. NLP Pipeline

C. Feature Extraction

Feature extraction is a step in the dimensionality reduction process, which divides and reduces a large set of raw data into smaller groupings. The fact that these enormous data sets have a large number of variables is the most crucial feature. To process these variables, a large amount of computational power is required. So, by selecting and merging variables into features, feature extraction aids in extracting the best feature from those large data sets, effectively lowering the amount of data. These features are simple to utilize while still characterizing the underlying data set precisely and uniquely. TFI-DF is utilized to extract features in this investigation. The TF-IDF (Term Frequency-Inverse Document Frequency) statistic is a numerical measure of how essential a term is to a document in a corpus or collection.

$$tf_{t,d}=\frac{n_{t,d}}{Number\ of\ terms\ in\ the\ document}$$

The IDF (Inverse Document Frequency) is a metric for determining the importance of a phrase. We need the IDF value since just computing the TF isn't enough to grasp the significance of words:

$$idf_t = log\frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

D. Implementation of Classifiers

In this section, proper parameters are used to train classifiers. For predictions, this framework used Logistic Regressor, SVM, and Naive Bayes models. SVM has a number of distinguishing characteristics, as a result of which it has gained notoriety and has shown promising experimental results. To partition the data points, SVM constructs a hyper level in authentic input space.

While Naive Bayes predicts the probability of different classes based on data, logistic regressors estimate probabilities using a logistic regression equation to determine the relationship between the dependent variable and one or more independent variables. The Random forest ensemble classifier was utilized as the classification algorithm, and it was built using a collection of tree-structured Classifiers.

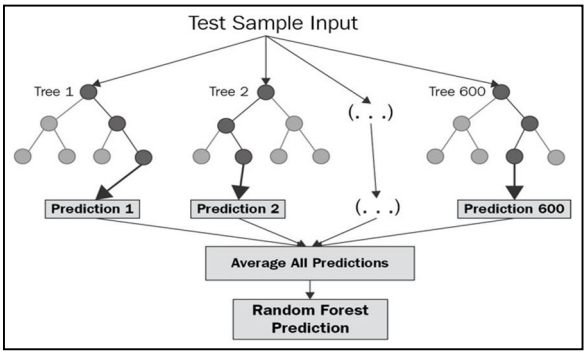


Fig.4. Random Forest classifier concept

The boosting is terminated on this random forest model, which was generated on 100 numbers of estimators. Following the construction of these classification models, the training dataset is utilised to make predictions, and then the performance is evaluated.

E. Performance Evaluation Metrics

When performing classification predictions, there are four types of outcomes that could occur: TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative). We have used four metrics for evaluating the performance of Fake Job detection system which are:

Accuracy: Accuracy is a measure that indicates the percentage of correct predictions made by our model..

Accuracy = TP+TN/TP+FP+FN+TN

Recall: Recall is the percentage of positives you properly identified out of all positives.

Recall = TP/TP+FN

Precision: The fraction of successfully identified positives out of all anticipated positives is known as precision.

Precision = TP/TP+FP

F1 Score: The harmonic mean of the model's precision and recall is what it's called.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

```
Classification Accuracy: 0.9778150633855331
Classification Report

              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5105
           1       1.00      0.54      0.70       259

    accuracy                           0.98      5364
   macro avg       0.99      0.77      0.85      5364
weighted avg       0.98      0.98      0.97      5364

Confusion Matrix

[[5105    0]
 [ 119  140]]
```

Fig.5. Classification Report

VI. EXPERIMENTAL RESULTS

Considering the fraud detecting problem, the situation of not detecting the job as fraud (low sensitivity) could be threatening for job-seekers. Whilst the low specificity (predicting legitimate job as fraud) may only cause a further inspection by a human given the fact that real jobs would be obvious to realize. However, the problem lies in tricking people with fraud jobs that may look like real ones. Table 6 shows a comparison between the proposed model.

Models Implemented	Accuracy
Logistic Regression	96%
Naive Bayes	84%
SVM	95%
Random Forest	97%

Fig.6. Accuracy of different Classifiers

VII. MODEL DEPLOYMENT

To make our model available for end users we are going to deploy our model using Python Flask on Heroku.

A. Flask

Flask is a Python-based web application framework. It features a number of modules that make it easy for a web developer to construct apps without having to worry about protocol management, thread management, and other such concerns.

B. Heroku

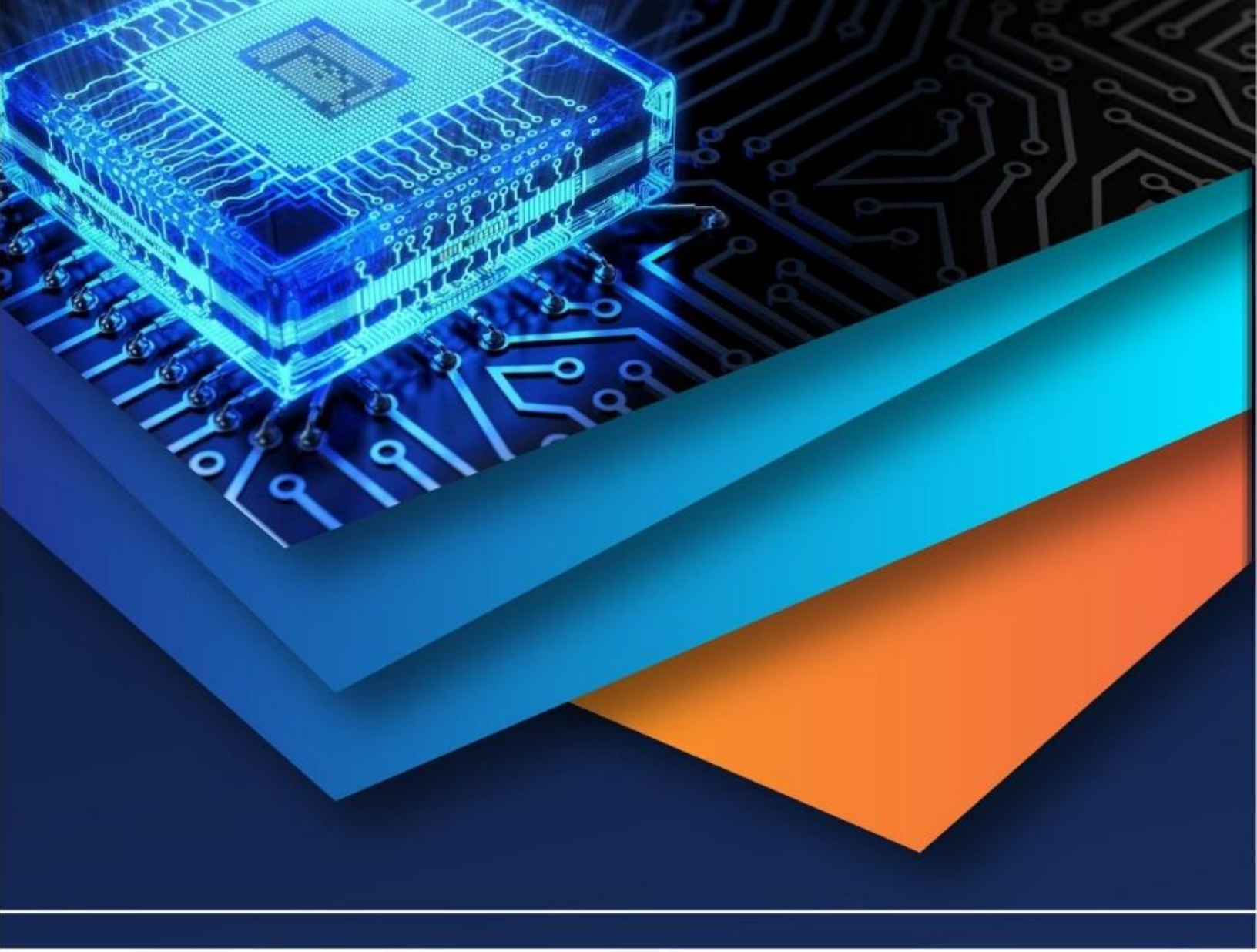
Heroku is a cloud platform that supports several programming languages in which we can deploy our applications.

VIII. CONCLUSIONS

Only reputable business offers will be sent to you. Several machine learning methods are proposed for detecting employment scams. In this work, we discuss counter measures. Supervised mechanism is used to demonstrate the utilization of many mechanisms. Classifiers for detecting job scams. The results of the experiments show that Random Forest is effective. The classifier exceeds its peers in classification. The proposed method had a 97 percent accuracy rate. Which is significantly greater than current approaches.

REFERENCES

[1] S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar, “Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms,” Rev. GEINTECGESTAO Inov. E Tecnol., vol. 11, no. 2, pp. 642–650, 2021.
[2] B. Alghamdi and F. Alharby, “An intelligent model for online recruitment fraud detection,” J. Inf. Secur., vol. 10, no. 03, p. 155, 2019.
[3] “Report | Cyber.gov.au.” https://www.cyber.gov.au/acsc/report (accessed Jun. 19, 2021).
[4] A. Pagotto, “Text Classification with Noisy Class Labels.” Carleton University, 2020.
[5] “Employment Scam Aegean Dataset.” http://emscad.samos.aegean.gr/ (accessed Jun. 19, 2021).
[6] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, “Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset,” Futur. Internet, vol. 9, no. 1, p. 6, 2017.
[7] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur and R. Mourya, "ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection," 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844879..
[8] Bandyopadhyay, Samir & Dutta, Shawni. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. International Journal of Engineering Trends and Technology. 68. 10.14445/22315381/IJETT-V68I4P209S.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY