iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61175

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 12 Issue IV Apr 2024- Available at www.ijraset.com

Web-based System for Identifying Fraudulent Job Advertisements with Machine Learning

Mrs. T. Kavitha[1] , K Hima Koushik Sriram[2], E V Rama Prasanth [3], B Siva Nagendra[4], Yalla Supraj Dev[5]
[1]*Assistant Professor* , [2, 3, 4, 5]*B.tech Students Department of Information Technology, Pragati Engineering College, Surampalem , Andhra Pradesh, India*

Abstract: ***To prevent false job advertisements on the internet, the study proposes a robotized framework that employs machine learning-based classification techniques. Various classifiers are employed to approve fraudulent web posts, and the results are compared to find the most effective job trick location technique. It helps identify fake job advertisements among a big number of posts. For detecting fake job postings, two major classifier types are considered: single classifiers and outfit classifiers. Nonetheless, the test results show that ensemble classifiers outperform single classifiers in task recognition.***
Keywords: Fake Job, Online Recruitment, Machine Learning, Ensemble Approach

I. INTRODUCTION

One of the most serious issues in the domain of Online Enrolment Fakes (ORF) has recently been addressed: business deception. Employers are increasingly choosing to post job openings online so that prospects can apply promptly and at the appropriate time. In any case, extortionists may be using this method on purpose, offering job seekers business in exchange for money. It is possible to abuse the legitimacy of fraudulent job announcements by posting them against a presumed organization. These phony job posting sites raise severe concerns about the need for a robotized system to detect fictitious employment opportunities and warn people in order to dissuade them from applying for such positions. As a result, a machine learning strategy is built that employs a few categorization computations to detect fraudulent messages. A classification device warns the customer by isolating fraudulent job postings from a broader collection of job advertisements. First, directed learning computations are considered as classification approaches to handle the challenge of finding tricks in job postings. A classifier considers information preparation while mapping an input variable to a target class. A basic example of applying the classifiers presented in the paper to distinguish between fake job postings and legitimate ones is provided. These classifier-based forecasts are roughly classified into two types: predictions based on a collection of classifiers and predictions based on a single classifier.

There is only one classifier for the prediction.

Classifiers are trained to forecast previously unknown test events. When recognizing fake job postings, the following classifiers are used:

1) *The Naive Bayes classifier:* is a supervised classification technique that employs the Bayes Theorem of Tacit Liability. Even though the classifier's probability forecasts are inaccurate, its judgment is usually correct in practice. In the script below, this classifier produces a very promising result: whether the traits are totally functionally dependent or independent. Because the independence hypothesis is necessary to predict the delicacy, the delicacy of this classifier is related to the quantum of information loss in the class rather than point dependences.
2) *Multi-Layer Perceptron Classifier:* With the training parameters changed, multi-layer perceptrons can be employed as supervised classification tools. A multilayer perceptron's number of nodes in each layer, as well as the number of hidden layers, can change depending on the task. The network design and training data are considered for determining the parameters.
3) *K-Nearest Neighbor Classifier:* K represents the nearest neighbor. Lazy learners, also known as classifiers, identify an object's identity in the feature space by comparing it to the most similar training samples. Before deciding on a class, the classifier considers the k closest items. The key challenge with this categorization strategy is determining a suitable value for k.
4) *Decision Tree Classifier:* A Decision Tree (DT) is a classification algorithm that employs tree-like structures. It gains classification expertise. A leaf node of DT is any target class; non-leaf nodes of DT serve as decision nodes to establish a specific test. The tests' outcomes are determined by either branch of that decision node. This tree grows from its root to a leaf node and eventually reaches that point. It is the technique for extracting classification output from a decision tree. Decision tree learning is one approach to spam filtering. This model can be setup and trained to forecast the target using predefined criteria.

II. LITERATURE SURVEY

"A Sensible Approach to Fraud in Online Recruitment" The purpose of this effort is to develop a reliable model for detecting fraud exposure in online recruitment environments, hence avoiding privacy violations and financial damages for individuals and organizations. This paper makes significant contributions by developing a trustworthy model for detecting online recruitment fraud (ORF) using an ensemble technique based on the Random Forest classifier. Compared to other types of electronic fraud detection, detecting online recruitment fraud is relatively new, with few studies conducted on the subject. To meet the study's objectives, the researcher proposed a detection model. For classification and detection, an ensemble classifier based on Random Forest is used, and feature selection is done using the support vector machine approach. The model is applied using a publicly accessible dataset called Employment Scam Aegean, also known as the EMSCAD dataset. Preprocessing was done initially, followed by selection and classification. The findings showed an accuracy of 97.41%. Furthermore, the findings identified the fundamental features and determining variables in recognition, such as having a firm profile, a logo, and an industry characteristic.

"A Comprehensive Analysis of the Naïve Bayes Classifier" Analyzing the naive Bayes classifier empirically

The naive Bayes classifier assumes feature independence within a class, which greatly simplifies learning. Despite the typical error of presuming independence, naïve Bayes often outperforms more advanced classifiers. Our primary goal is to understand the properties of the data that influence naive Bayes performance. Using Monte Carlo simulations, our method allows for a rigorous comparison of categorization performance across multiple classes of randomly generated issues. According to our study of the association between distribution entropy and classification error, low-entropy feature distributions produce good naive Bayes performance. Furthermore, we demonstrate that naïve Bayes works best in two distinct scenarios: completely independent features (as predicted) and functionally dependent features. This is because naive Bayes works effectively with some nearly-functional feature dependencies.

This is astonishing. Another surprise finding is that the degree of feature dependencies, measured as class-conditional mutual information across features, does not immediately correspond with naive Bayes performance. Alternatively, the amount of class knowledge lost due to the independence assumption is a more trustworthy predictor of naive Bayes correctness.

"The Analysis of Binomial Random Variables and Bailey's Theorem" The author provides a reasonably useful application of the Bayes theorem to the study of random variables with binomial distributions. The technique's reliability for one or two random variables has been established in previous publications (Walters, 1985; Walters, 1986a), and an extension of the concept to multiple random variables is presented. The method is illustrated with two biometric instances.

"Classification and regression using multilayer perceptrons" We discuss the theory and use of the multilayer perceptron. In order to apply this method to real-world problems, we need to investigate a variety of relevant concerns. A variety of examples are used to compare the multilayer perceptron to other typical techniques. Regression and classification are two application fields that are discussed specifically. Implementation difficulties are addressed, including dynamics, architecture, and multilayer perceptron properties. Recent research is cited, particularly in the areas of function mapping and discriminant analysis.

"A Review of Decision Tree Classification Algorithms in Data Mining" The volume of data in the information sector is always growing due to advances in computer and computer network technologies. Examining this massive amount of data and drawing relevant conclusions from it is critical. Data mining is the process of extracting useful information from large quantities of inaccurate, noisy, confusing, and unpredictable data. The decision tree classification approach is a popular data mining methodology. The divide and conquer method is used in decision trees as a fundamental learning mechanism. A decision tree is a structure made up of root nodes, branches, and leaf nodes. Every leaf node has the class label, every branch displays the result of a test, and each internal node represents an attribute test. The root node is at the top of the tree. This paper focuses on the characteristics, challenges, advantages, and downsides of various decision tree algorithms (ID3, C4.5, and CART).

III. SYSTEM ANALYSIS

A. Existing System

Several studies have found that the detection of email spam, review spam, and fake news has received significant attention in the field of online fraud detection.

1) Check Spam Identification

People commonly share their thoughts on the things they buy in internet forums. It might help other customers choose products. Techniques for detecting these reviews must be developed, as spammers can now manipulate reviews to increase their revenue. This can be accomplished by extracting features from reviews using Natural Language Processing (NLP).

Following that, machine learning techniques are used to these features. Lexicon-based strategies may be a viable alternative to machine learning techniques that employ a corpus or vocabulary to eliminate spam reviews.

2) Email Spam Detection

User mailboxes are frequently inundated with unsolicited bulk emails, generally known as spam emails. This can result in the consumption of bandwidth and an unavoidable storage problem. Gmail, Yahoo Mail, and Outlook service providers use Neural Network-based spam filters to address this issue. We compare adaptive spam filtering, content-based filtering, case-based filtering, heuristic-based filtering, memory or instance-based filtering, and heuristic-based filtering for detecting email spam.

Fake News Detection

Echo chamber effects and hostile user profiles are common features of fake news on social media. The three essential views on fake news detection are the development of fake news, its distribution, and the interaction of a user with fake news. Following the recovery of social context and content-related data, machine learning models are used to detect fake news.

DISADVANTAGES OF THE EXISTING SYSTEM

If the dataset used to train the machine learning models is unequal, with fewer positive (fraudulent) cases than negative (genuine job adverts), the models may be unable to generalize to new, unknown data effectively.

a) *Challenges with Feature Engineering:* Creating features for the machine learning model that accurately describe job adverts might be difficult. If key features are absent or poorly represented, the model's performance may deteriorate.

b) *Adaptability to Changing Scams:* As fraudulent activities evolve over time, new strategies may emerge. The incapacity of the current system to quickly adapt to new scam types may result in false negatives.

c) *Explainability and Interpretability:* Transparency and interpretability may be absent in some machine learning models, particularly complicated ones such as ensemble classification. It is crucial to understand the thinking behind a model's output, especially in sensitive areas such as fraud detection.

d) *Scalability:* When faced with a large number of job advertisements, the existing system's performance may decrease. Scalability concerns may develop if the system is not designed to manage large amounts of data adequately. Dependency on training data: The representativeness and quality of the training data have a significant impact on the efficacy of machine learning algorithms. If the training data fails to capture the full range of bogus job advertisements, the model's practical performance may suffer.

e) *Complex machine learning models:* particularly ensemble classifiers, may necessitate large processing resources for training and inference. There may be hardware and time constraints.

f) *False Positives:* The system may misidentify real job posts as fraudulent in order to generate false positives. This may upset users and lead to a loss of faith in the system. Regulatory Compliance: Using machine learning models for fraud detection may pose ethical and legal concerns, depending on the application domain. Ensuring compliance with relevant rules is critical.

B. Proposed System

The proposed strategy uses cutting-edge machine learning techniques to assist job placement sites in identifying and reducing bogus job postings. Using cutting-edge classification algorithms, the system improves feature engineering, boosts resilience to evolving fraud, and ensures excellent scalability for managing massive job posting volumes. attempts to overcome the disadvantages of this method. Furthermore, the proposed approach prioritizes interpretability and explainability to provide a better understanding of the reasoning process involved in fraud detection. Based on experimental data, the system considers the advantages of ensemble classifiers over individual classifiers in order to get optimal results. The ultimate goal is to provide a robust and dependable solution that lowers false positives and successfully detects bogus job advertising, hence increasing user trust in recruiting websites. is. The proposed method also addresses ethical and legal concerns about the use of machine learning models in fraud detection applications, as well as ensuring regulatory compliance. With these improvements, it is expected that the proposed system will set a new standard for detecting fraudulent job applications and facilitating more safe and trustworthy online job searches.

ADVANTAGES OF THE PROPOSED SYSTEM

1) *Enhanced Fraud Detection Accuracy:* Using complex machine learning techniques such as ensemble classifiers, the proposed method dramatically enhances the accuracy of recognizing phony job postings.

By using new algorithms, the system can more reliably identify trends and abnormalities associated with fraud, improving its ability to identify phony job adverts.

2) *Adaptability to Emerging Scams:* Unlike existing solutions, the proposed one is designed to respond fast to changing fraud methods. Because of continual learning and improvements, the system can successfully detect and combat new sorts of fraudulent activity in the dynamic domain of online job recruitment.
3) *Improved Scalability:* The system's scalability allows for the efficient processing and analysis of a large number of job ads. This benefits job recruitment systems with high user involvement levels in particular by allowing the system to handle larger data volumes while maintaining efficacy.
4) *Enhanced Explainability and Interpretability:* The suggested approach prioritizes interpretability and transparency, providing detailed insights into how machine learning models generate decisions. This function not only fosters system trust, but it also allows users and platform managers to understand why a job posting has been flagged as potentially fake.
5) *Decreased False Positives:* The suggested approach intends to reduce false positives by optimizing machine learning models and feature engineering strategies. This is required to maintain the job recruitment platform's outstanding user experience, avoid incorrectly classifying genuine job ads as phony, and foster enhanced user faith in the system's legitimacy.

IV. SYSTEM DESIGN

SYSTEM ARCHITECTURE

Below diagram depicts the whole system architecture.

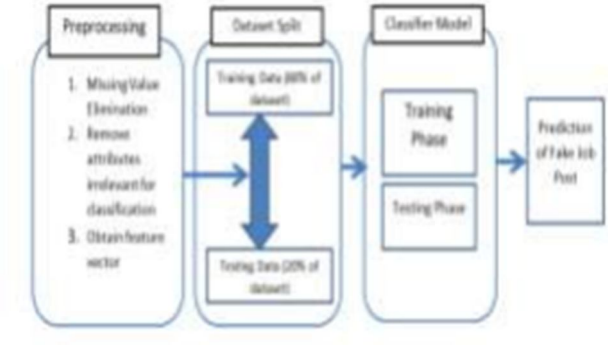


Fig 1. System Architecture

V. SYSTEM IMPLEMENTATION

MODULES

1) *Module for Data Preprocessing:* This module cleans and prepares data for analysis. It includes duties such as filling in missing figures, removing extraneous information, and standardizing data formats. Data preparation ensures that the input data is of high quality and consistent for future machine learning model training.
2) *Module for Feature Engineering:* Increasing the effectiveness of machine learning models necessitates feature engineering. This module selects and modifies essential properties from the dataset in order to offer meaningful input to the classification algorithms. Text analysis and the extraction of significant job-related attributes yield a feature set that includes the fundamental qualities of job postings.
3) *The Machine Learning Classification Module:* trains and deploys machine learning classifiers. It assesses feature-rich data and uses individual and group classifiers to estimate the validity of job advertisements. This section covers the steps for selecting classifiers, training models, and optimizing them.
4) *Model Evaluation and Comparison Module:* After training, this module examines the classifiers' performance using metrics such as accuracy, precision, recall, and F1-score. It also allows you to compare and contrast different classifiers to see which model is better at detecting fake job listings. Model evaluation is critical for determining the best performance strategy and fine-tuning parameters.
5) *User Interface and Reporting Module:* The major purpose of this module is to create an intuitive user interface for system interaction. Users can submit job listings for analysis and view the findings using the tools provided.

6) The module also generates detailed reports on the classification results, indicating whether a job posting is marked as potentially fraudulent or legitimate. Simple and basic visuals may be provided to help users understand the system's conclusions.

VI. EXPERIMENTAL RESULTS

A dataset including both false and genuine job advertisements is utilized to train and assess all of the aforementioned classifiers for detecting fraudulent job listings. Table 2 shows the results for classifiers that use ensemble approaches, and the next table compares the classifiers based on metrics. Figure 2 shows the accuracy, f1-score, Cohen-kappa score, and MSE for all classifiers combined.

Used	Accuracy	Precision Score	Recall Score	F1 Score
Logistic Regression	97.50186428038778	71.32670553700844<	96.78476492908649<	78.92893996247655<
Decision Tree	97.81879194630872	89.17677658586449<	85.46277665995976<	87.21626897354007<
Naive Bayes	95.76808351976138	50.0<	47.88404175988069<	48.9191505570898<
Random Forest	98.37807606263982	81.67912844449742<	97.85301981429282<	87.97983193277311<

fig 2. performance comparison chart for ensemble classifier-based prediction

Fig 3. Result For predicting Fake job posts

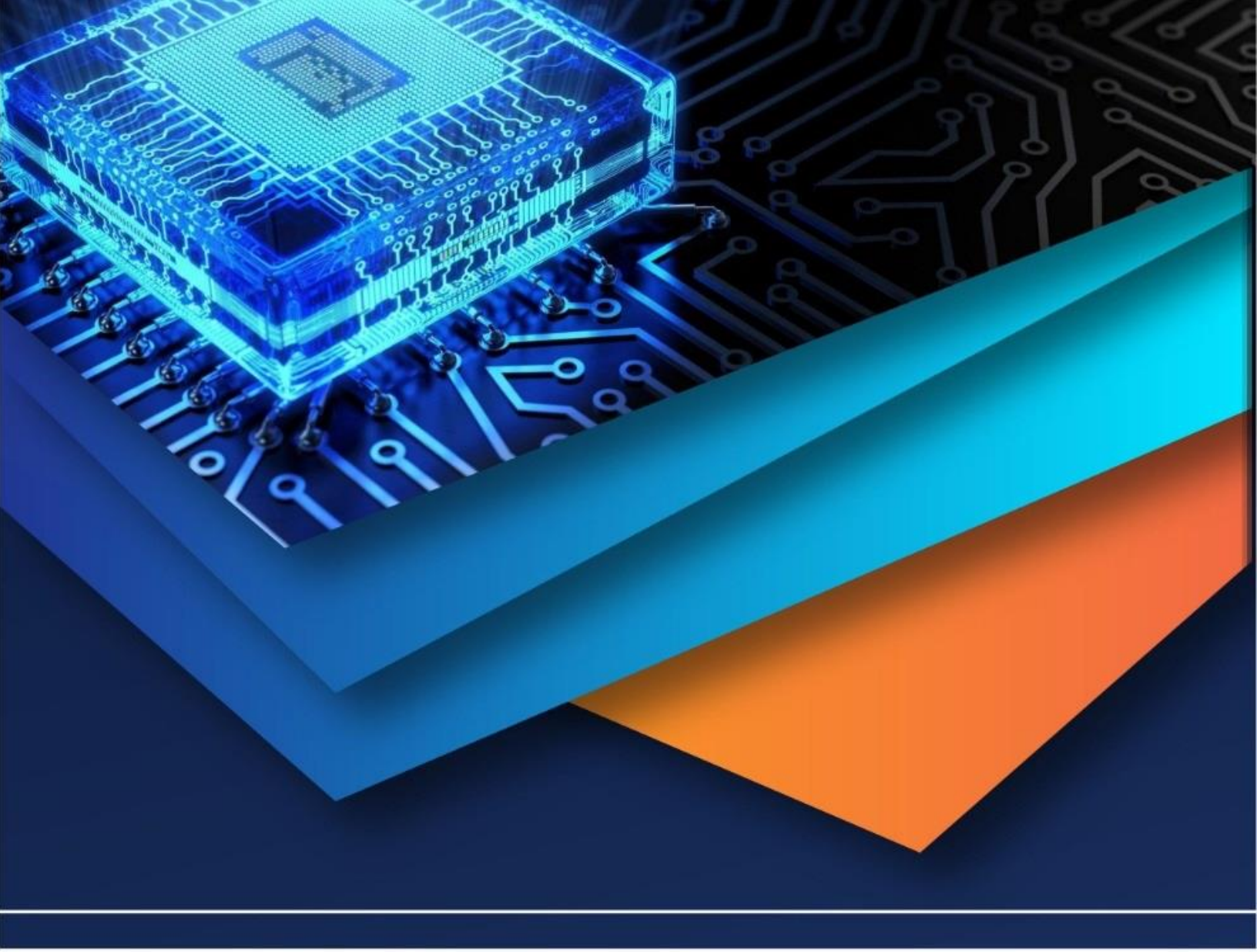
VII. CONCLUSION AND FUTURE WORK

Recognizing employment scams could help job seekers receive only authentic job offers from employers. Several machine learning techniques are proposed in this paper as countermeasures for detecting employment scams. The supervised technique is used to demonstrate the use of multiple classifiers for detecting employment fraud. Experiments show that the Random Forest classifier outperforms its peer classification technique. The suggested strategy has an accuracy of 98.27%, which is a significant improvement above current techniques.

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
[9] Andrii Shalaginov, Jan William Johnsen, Katrin Franke, "Cyber Crime Investigations in the Era of Big Data", IEEE International Conference on Big Data, 2017, pp. 3672-3676.
[10] Sokratis Vidros, Constantinos Kolias, Georgios Kambourakis and Leman Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset", Future Internet 2017, pp. 2-19.
[11] Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake news detection on social media: A data mining perspective." ACM SIGKDD explorations newsletter 19, no. 1 (2017): 22-36.
[12] Devsmit Ranparia; Shaily Kumari; Ashish Sahani, "Fake Job Prediction using Sequential Network", IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp.339-343
[13] Syed Mahbub, Eric Pardede, "Using Contextual Features for Online Recruitment Fraud Detection", 27th International Conference on Information Systems Development, 2018.
[14] Najma Imtiaz Ali, Suhaila Samsuri, Muhamad Sadry, Imtiaz Ali Brohi, Asadullah Shah, "Online Shopping Satisfaction in Malaysia: A Framework for Security, Trust and Cybercrime", 6th International Conference on Information and Communication Technology for The Muslim World, 2016, pp. 194-198.
[15] Vidros, Sokratis; Kolias, Constantinos; Kambourakis, Georgios, "Online recruitment services: another playground for fraudsters", Computer Fraud & Security, 2016, pp. 8-13.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY