Machine Learning-Powered Web Application for Predicting and Identifying Fake Job Listing

KANAKALA SS PRAVEEN KUMAR
Department of Bachelor of computer applications Nri institute , nagarabhavi 2nd stage, bangalore
Email: praveen.kanakala@gmail.com
Dr P VAMSI KRISHNA RAJA
Professor, Department of Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Seetha Ramapuram
Mail: drpvkraja@ieee.org
B.R. AMBEDKAR KOTA
Lecturer in computer science SKVT Government Degree College, Rajamahendravaram

Abstract: This paper proposes an automated system that leverages machine learning-based classification methods to identify fraudulent job postings on the web. Various classifiers are employed to validate the fake posts, and the outcomes are compared to determine the most effective job scam detection strategy. The system aids in discerning fake job advertisements among numerous posts. Two primary types of classifiers are utilized: single classifiers and ensemble classifiers. The test results indicate that ensemble classifiers outperform single classifiers in detecting scams.

Keyword: Machine learning, Discerning fake job Advertisements, Single classifiers and ensemble classifiers detecting scams

I. INTRODUCTION

Job scams have become a significant issue in the domain of Online Recruitment Frauds (ORF). Recently, many companies prefer to post their job openings online for easy and timely access by job-seekers. However, this convenience is often exploited by fraudsters who post fake job ads to scam job-seekers, usually involving monetary transactions. These fraudulent job ads can also misuse the reputation of reputable companies. Therefore, an automated tool to detect and report fake job listings is essential to prevent job-seekers from falling victim to such scams. This research employs machine learning techniques to classify job postings as fake or legitimate. The classifiers used for prediction are categorized into single classifiers and ensemble classifiers.

Single Classifier Based Prediction:

- **Naive Bayes Classifier:** The Naive Bayes classifier, a supervised learning algorithm, utilizes Bayes Theorem of conditional probability. Despite potential inaccuracies in its probability estimates, it performs effectively, particularly when feature independence or complete dependence is assumed. The classifier's accuracy is influenced more by the loss of class information due to the independence assumption rather than feature dependencies.
- **Multi-Layer Perceptron Classifier:** Multi-layer perceptrons (MLPs) serve as supervised classification tools when training parameters are fine-tuned. The structure, including the number of hidden layers and nodes, varies depending on the problem, training data, and network design.
- **K-Nearest Neighbor Classifier:** Known as lazy learners, K-Nearest Neighbor (K-NN) classifiers identify objects based on proximity to training samples in the feature space. The classifier determines the class by evaluating the closest k objects, with the value of k being crucial for classification accuracy.
- **Decision Tree Classifier:** Decision Trees (DT) classify data using tree-like structures, where each leaf node represents a class, and non-leaf nodes act as decision nodes. The tree progresses from the root to leaf nodes, using branches to represent test outcomes. DT learning, utilized in spam filtering, predicts targets based on specific criteria through a trained model.

II. LITERATURE SURVEY

TITLE: "An Advanced Model for Detecting Fraud in Online Recruitment"

This study seeks to safeguard individuals and organizations from privacy breaches and financial losses by creating a dependable

Manuscript received February 10, 2024; Revised February 25, 2024; Accepted March 12, 2024

model for identifying fraudulent activities in online recruitment environments. This research makes a notable contribution by developing a robust detection model for online recruitment fraud (ORF) utilizing an ensemble approach based on the Random Forest classifier. Unlike other forms of electronic fraud detection, online recruitment fraud is relatively new and less studied. The proposed detection model addresses this gap. Feature selection is performed using the support vector machine approach, while classification and detection are carried out with an ensemble classifier based on Random Forest. The model is tested using the publicly available Employment Scam Aegean dataset (EMSCAD), following a pre-processing step. The results demonstrated an accuracy of 97.41%. Key features and variables identified include company biography, corporate logo, and industry details.

TITLE: "A Comprehensive Study of the Naïve Bayes Classifier"

The Naive Bayes classifier simplifies the learning process by assuming that features within a class are independent. Despite the fact that this assumption is often unrealistic, Naive Bayes frequently outperforms more sophisticated classifiers. This study aims to understand the data characteristics that affect Naive Bayes performance. Using Monte Carlo simulations, we systematically explore classification performance across various classes of randomly generated problems. We examine the impact of distribution entropy on classification error and find that low-entropy feature distributions yield good Naive Bayes performance. Additionally, we show that Naive Bayes performs well with nearly-functional feature dependencies, achieving optimal performance in cases of either fully independent features or completely dependent features. Surprisingly, Naive Bayes accuracy is not directly related to the degree of feature dependencies but rather to the amount of class information lost due to the independence assumption.

TITLE: "Application of Bayes's Theorem to Binomial Random Variables"

This paper demonstrates a practical application of Bayes' theorem to the analysis of binomial random variables. Previous works by Walters (1985, 1986a) have shown the method's reliability for one or two random variables. This study extends the approach to multiple random variables, providing two biometric examples to illustrate the method.

TITLE: "Multilayer Perceptrons for Classification and Regression"

This work discusses the theory and application of multilayer perceptrons (MLPs). We address various issues relevant to applying this method to real-world problems, providing several examples to demonstrate how MLPs compare with conventional methodologies. The focus is on the use of MLPs in classification and regression, addressing implementation questions such as MLP architecture, dynamics, and associated features. Recent advancements, particularly in discriminant analysis and function mapping, are also discussed.

TITLE: "A Survey of Decision Tree Algorithms for Classification in Data Mining"

With the advancement of computer and network technology, the volume of data in the information sector continues to grow. It is crucial to analyze this vast amount of data and extract relevant information. Data mining involves extracting meaningful knowledge from large sets of incomplete, noisy, fuzzy, and random data. Decision tree classification is one of the most common data mining techniques, using the divide and conquer method as a fundamental learning mechanism. A decision tree consists of a root node, branches, and leaf nodes. Each internal node represents a test on an attribute, each branch represents the test's outcome, and each leaf node contains a class label. This study examines the features, challenges, advantages, and drawbacks of various decision tree algorithms (ID3, C4.5, and CART).

TITLE: "Machine Learning for Email Spam Filtering: Review, Approaches, and Open Research Problems"

The need to develop robust and reliable antispam filters has increased with the rise of unsolicited emails, known as spam. Recent advancements in machine learning have enabled effective identification and filtering of spam emails. This review provides an in-depth analysis of several widely used machine learning-based email spam filtering techniques, covering key concepts, methods, effectiveness, and current research directions. Initially, the study examines how top internet service providers (ISPs) like Gmail, Yahoo, and Outlook apply machine learning techniques to spam filtering. It discusses the general process of email spam filtering and various machine learning methods used. The review contrasts the advantages and drawbacks of current techniques and addresses unresolved issues in spam screening. The study proposes deep learning and adversarial learning as future approaches to effectively tackle spam email challenges.

III. SYSTEM ANALYSIS

A. EXISTING SYSTEM

Numerous studies highlight that the identification of fake news, email spam, and review spam has been a focal point in online fraud detection.

1. **Review Spam Detection:** Individuals often share their opinions about purchased products on internet forums, which can aid other buyers in making decisions. However, spammers can manipulate these reviews to increase their profits, necessitating methods to identify such fake reviews. This can be achieved through Natural Language Processing (NLP) to extract features from reviews, followed by applying machine learning techniques. Alternatively, lexicon-based methods using a corpus or dictionary can also be employed to eliminate spam reviews.
2. **Email Spam Detection:** Bulk unsolicited emails, known as spam, frequently fill user inboxes, leading to bandwidth usage and storage issues. Neural Network-based spam filters are used by providers like Gmail, Yahoo Mail, and Outlook to tackle this problem. Various approaches to email spam detection include adaptive spam filtering, content-based filtering, case-based filtering, heuristic-based filtering, and memory or instance-based filtering.
3. **Fake News Detection:** Fake news on social media is characterized by echo chamber effects and malicious user profiles. Key research perspectives in fake news detection include the creation of fake news, its dissemination, and the user-news relationship. To detect fake news, features related to social context and news content are extracted and analyzed using machine learning models.

DISADVANTAGES OF THE EXISTING SYSTEM

1. **Generalization Issues:** Machine learning models may not effectively generalize to new, unseen data if the training dataset is imbalanced, with fewer positive (fraudulent) cases compared to negative (legitimate job listings) ones.
2. **Feature Engineering Challenges:** Designing features that accurately describe job ads can be difficult. Missing or inadequately represented features can hamper model performance.
3. **Adaptability to Changing Scams:** Fraudulent tactics evolve over time. The current system may struggle to adapt quickly to new scam types, potentially resulting in false negatives.
4. **Explainability and Interpretability:** Complex machine learning models, such as ensemble classifiers, may lack transparency and interpretability. Understanding why a model makes a specific prediction is crucial, especially in sensitive areas like fraud detection.
5. **Scalability:** The performance of existing systems may decline when handling a large number of job listings. Scalability issues can arise if the system is not designed to manage large data volumes efficiently.
6. **Dependency on Training Data:** The effectiveness of machine learning models heavily depends on the quality and representativeness of the training data. If the training data does not fully capture the range of fraudulent job ads, the model may underperform in real-world situations.
7. **Computational Resources:** Training and inference with complex machine learning models, especially ensemble classifiers, require significant computational resources, which can be a constraint in terms of time and hardware.
8. **False Positives:** The system may incorrectly classify legitimate job listings as fraudulent, leading to user frustration and loss of trust in the system.
9. **Regulatory Compliance:** There may be legal and ethical concerns associated with using machine learning models for fraud detection. Ensuring compliance with relevant regulations is essential.

B. PROPOSED SYSTEM

The proposed system aims to enhance the detection and mitigation of fake job postings on job placement platforms by integrating advanced machine learning techniques. It employs state-of-the-art classification algorithms, improves adaptability to evolving fraud tactics, refines the feature engineering process, and ensures superior scalability to handle large volumes of job postings efficiently. Additionally, the system emphasizes explainability and interpretability to provide clear insights into the fraud detection decision-making process. To achieve optimal results, the system leverages the advantages of ensemble classifiers over individual classifiers, as evidenced by experimental results. The goal is to offer a robust and reliable tool that not only effectively identifies fraudulent job advertisements but also minimizes false positives, thereby increasing user trust in recruitment platforms. Moreover, the system aims to maintain regulatory compliance and address legal and ethical issues related to the use of machine learning models in fraud detection applications. With these improvements, the proposed system is expected to set a new standard in detecting fake job applications and contribute to a safer and more reliable online job search experience.

ADVANTAGES OF THE PROPOSED SYSTEM

1. **Enhanced Fraud Detection Accuracy:** The proposed system utilizes advanced machine learning techniques, such as ensemble classifiers, to significantly improve the accuracy of detecting fake job postings. By leveraging complex algorithms, the system can better identify patterns and anomalies associated with fraud,

leading to higher precision in recognizing fake job advertisements.

2. **Adaptability to Emerging Scams:** Unlike existing systems, the proposed solution is designed to dynamically respond to evolving fraud tactics. Through continuous learning and updates, the system can effectively identify and counter new types of fraudulent activities in the ever-changing landscape of online job recruitment.
3. **Improved Scalability:** The system is built for scalability, enabling efficient processing and analysis of large volumes of job postings. This is particularly beneficial for job recruitment platforms with high user engagement, allowing the system to handle increased data volumes without compromising performance.
4. **Enhanced Explainability and Interpretability:** The proposed approach prioritizes transparency and interpretability, offering clear insights into the decision-making process of machine learning models. This feature not only helps build trust in the system but also allows users and platform administrators to understand why certain job postings are labeled as potentially fraudulent.
5. **Reduced False Positives:** The system aims to minimize false positives by fine-tuning machine learning models and feature engineering techniques. This is crucial for maintaining a positive user experience on the job recruitment platform, preventing legitimate job ads from being wrongly classified as fraudulent, and fostering increased user confidence in the system's accuracy.

IV.SYSTEM DESIGN

SYSTEM ARCHITECTURE

Below diagram depicts the whole system architecture.

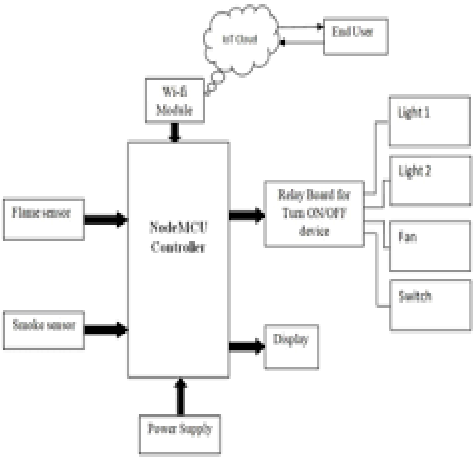


Fig 1. System Architecture

V. SYSTEM IMPLEMENTATION

MODULES

Data Preprocessing Module:

This module cleans and prepares raw data for analysis. It entails duties including resolving missing numbers, deleting extraneous information, and standardizing data formats. Data preprocessing ensures the quality and consistency of the input data for future machine learning model training.

Feature Engineering Module:

Feature engineering is a critical step toward improving the effectiveness of machine learning models. This module consists of choosing and manipulating significant features from the dataset in order to offer meaningful input to the classifiers. Text analysis and the extraction of key job-related attributes are used to generate a feature set that encapsulates the essential characteristics of job posts.

Machine Learning Classification Module:

This fundamental module trains and deploys machine learning classifiers. It uses both single and ensemble classifiers to assess feature-rich data and estimate the validity of job advertisements. This section includes the process of selecting classifiers, training models, and optimizing them.

Model Evaluation and Comparison Module:

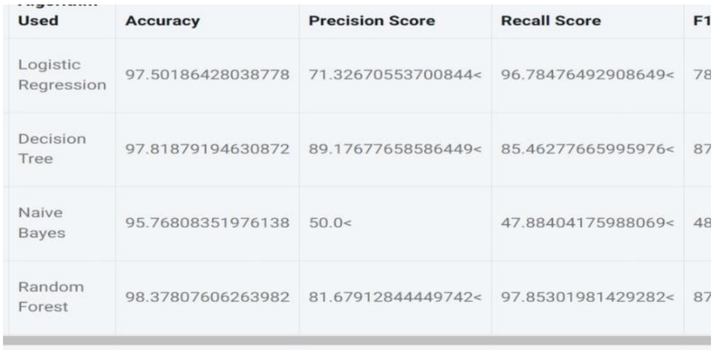
After training the classifiers, this module assesses their performance using metrics including accuracy, precision, recall, and F1-score. It also allows for a comparison analysis of various classifiers to choose the most effective model for detecting fake job postings. Model evaluation is crucial for fine-tuning parameters and choosing the best-performing method.

User Interface and Reporting Module:

This module focuses on creating a user-friendly interface for interacting with the system. It has tools that allow users to submit job listings for analysis and examine the outcomes. Furthermore, the module offers thorough reports on categorization results, indicating if a job ad is tagged as potentially fake or authentic. Clear and intuitive visualizations may also be included to assist user understanding of the system's findings.

V. EXPERIMENTAL RESULTS

All of the above-mentioned classifiers are trained and tested to detect bogus job posts using a dataset that includes both false and authentic posts. The next table compares the classifiers in terms of assessing metrics, and Table 2 shows the results for classifiers that use ensemble approaches. Figure 2 shows the overall performance of all classifiers in terms of accuracy, f1-score, Cohen-kappa score, and MSE.



Used	Accuracy	Precision Score	Recall Score	F1
Logistic Regression	97.50186428038778	71.32670553700844<	96.78476492908649<	78
Decision Tree	97.81879194630872	89.17677658586449<	85.46277665995976<	87
Naive Bayes	95.76808351976138	50.0<	47.88404175988069<	48
Random Forest	98.37807606263982	81.67912844449742<	97.85301981429282<	87

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

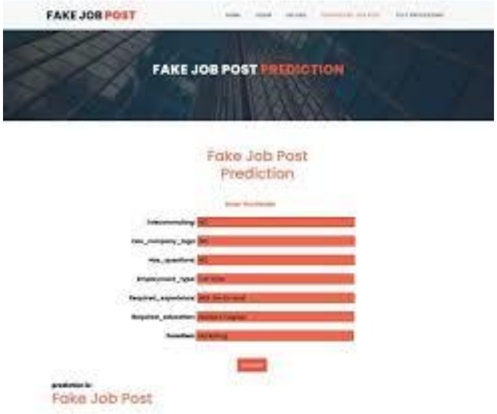


Fig 3. Result For predicting Fake job posts

VI . CONCLUSION AND FUTURE WORK

Employment scam identification can help job searchers acquire only authentic offers from employers. This research proposes numerous machine learning methods as countermeasures to detect employment scams. The supervised technique is used to demonstrate the utilization of several classifiers for job fraud detection. Experimental data show that the Random Forest classifier outperforms its peer classification technology. The proposed strategy obtained accuracy of 98.27%, which is significantly higher than the existing methods.

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.

Author Profile

KANAKALA SS PRAVEEN KUMAR
Department of Bachelor of computer applications
Nri institute , nagarabhavi 2nd stage, bangalore
Email: praveen.kanakala@gmail.com

Dr P VAMSI KRISHNA RAJA
Professor, Department of Computer Science and Engineering, Swarnandhra College of Engineering and Technology, Seetha Ramapuram
Mail: drpvkraja@ieee.org

B.R. AMBEDKAR KOTA
Lecturer in computer science SKVT Government Degree College,
Rajamahendravaram