Fake Job Post Prediction Through a Comparative Study on Diverse Data Mining Techniques

UTHARA SUDHIR[1], DR. V. UMARANI[2]
[1] *CNIS Student, Computer Networks and Information Security, Department of IT, Jawaharlal Nehru Technological University Hyderabad, Kukatpally, Hyderabad*
[2] *Professor, Computer Networks and Information Security, Department of IT, Jawaharlal Nehru Technological University Hyderabad, Kukatpally, Hyderabad*

Abstract*— In the contemporary digital landscape, the surge of deceptive job postings presents a substantial challenge for both job seekers and employers. This paper aims to tackle this pressing issue by developing sophisticated predictive models. The objective is to identify patterns distinguishing genuine job postings from fraudulent ones. By leveraging the Employment Scam Aegean Dataset (EMSCAD) from Kaggle, which includes both real and fake job postings, our study employs various data mining techniques and classification algorithms such as Decision Tree, Random Forest, Logistic Regression, K-NN, Support Vector Machine, Naïve Bayes, and Neural Network. Through rigorous testing and comparative analysis of various methodologies, the study seeks to establish the most effective approach for predicting and monitoring fraudulent job postings, thereby contributing to a safer and more reliable job market environment. Enhancing the precision of fraud detection not only safeguards job seekers but also fortifies the integrity of the job market.

Index Terms- Data Mining, Fake Job Post, Machine Learning, Prediction, Python, Random Forest, Regression

I. INTRODUCTION

Data Mining and Machine Learning represent pivotal and burgeoning fields within AI, employing diverse analytical techniques and statistical methodologies to enable systems to learn and perform based on historical data. These disciplines harness machines' capacity to assimilate knowledge from vast datasets, thereby improving performance through experiential learning. By leveraging past experiences, these fields empower the development of intelligent systems capable of addressing specific challenges.

The motivation to work on this study is to understand the in-depth applications of the Machine Learning models and Data Mining techniques. These AI-driven approaches are instrumental in tackling complex tasks by uncovering patterns, trends, and insights from extensive datasets. This capability enables businesses and researchers to extract valuable information, make informed decisions, and innovate across various domains. The study will specifically employ the Employment Scam Aegean Dataset (EMSCAD), comprising 18,000 samples on which various techniques are worked upon to extract different features and patterns.

Several algorithms namely, Decision Tree, Random Forest, Logistic Regression, Naïve Baye's model, K-NN, Neural Network, Support Vector Machine and subjected them to the dataset for processing. After evaluating the accuracy of various models, the output is generated from the model demonstrating the highest accuracy. Python was selected as the programming language for our project due to its exceptional efficiency in developing applications, supported by its extensive array of libraries and functions. The model codes are implemented in Python files, and Python itself is employed for conducting comparative analyses. Numerous academic papers and research studies have focused on pioneering techniques, including the development of hybrid models combining different methodologies. Python's versatility and robust ecosystem have facilitated the seamless integration of these advanced techniques into our project. This language's scalability and community support have further ensured that we can continuously refine and enhance our models based on the latest research findings and methodologies.

This paper demonstrates the outlined patterns, systematic processes, the overview of existing models, the proposed model's description, the analysis of

models, the conclusion and the future scope of the project.

II. RELATED WORK

There are many works related to the literature of the subjected topic. The outline of few of them is described below.

- Online Recruitment Frauds (ORF) pose a significant threat due to the widespread adoption of digital hiring processes, which aim to simplify and optimize recruitment. However, this digital exposure has also amplified risks such as privacy breaches for applicants and employees. The most prevalent ORF, employment scams, can lead to severe consequences like financial loss and identity theft. Organizations are increasingly investing in technologies to swiftly detect and remove fraudulent job postings, safeguarding job seekers and preserving the integrity of recruitment practices[1].
- Detecting fake reviews is critical amidst the proliferation of online platforms where reviews influence decisions and innovations. Despite years of research using supervised learning, online review systems are vulnerable to spam and manipulation, especially in the context of deceptive job postings that lure applicants with false promises of high-paying remote positions. Ensuring the authenticity of job listings is essential to protect individuals from divulging personal information or falling prey to upfront fee scams[1].
- Advanced models are being developed to combat Online Recruitment Fraud (ORF), aiming to protect individuals and organizations from financial losses and privacy breaches. Detection techniques utilize data mining, natural language processing (NLP), and machine learning algorithms to analyse job descriptions, company profiles, and applicant behaviours for inconsistencies indicative of fraudulent activities. Community feedback mechanisms further enhance detection accuracy by allowing users to report suspicious job postings promptly[4].
- Predicting suitable job matches using machine learning models helps employers assess candidates' qualifications and skills accurately. Implementing robust detection systems not only safeguards job seekers from financial harm but also fosters trust in online job platforms. Ensuring transparency and fairness in recruitment processes is crucial for maintaining a healthy job market ecosystem where genuine opportunities thrive and fraudulent activities are mitigated effectively.

Also, it is now, the most needed study to ensure the authenticity and integrity of the job market are precise.

III. EXISTING METHODOLOGY

Several studies on identifying fraudulent job postings focus on detecting deceptive accounts, known as content polluters or spammers. Researchers analyse user demographics, social network structures, posted content, and temporal activity patterns to differentiate legitimate users from malicious ones.

Demographic insights highlight differences in profile completeness, consistency of information, and interaction histories, revealing potential discrepancies indicative of suspicious behaviour. Social graph analysis examines follower/following patterns and network formations used to amplify false information, exposing fake accounts[6,7].

Content analysis evaluates post relevance, coherence, and frequency, identifying repetitive or misleading content dissemination patterns. Temporal behaviour analysis scrutinizes posting frequency and timing, detecting irregularities that suggest automated or coordinated efforts to spread deceptive content.

Integrating these analyses enhances detection capabilities, empowering researchers and platform developers to mitigate the impact of fake accounts online. This proactive approach safeguards users from misinformation and fraud, reinforcing platform integrity.

Gao et al. applied clustering techniques to identify fake posts by analysing text and URL similarities, uncovering clusters with frequent posting bursts. Their incremental clustering method dynamically updates to capture evolving deceptive patterns in social media campaigns, proving effective in combating online misinformation[1].

IV. SYSTEM CONCEPTION

There are several crucial steps to ensure the reliability and efficacy of models and findings. The following steps demonstrate the conception adopted for this study.

Dataset Collection: The initial phase involves gathering relevant data from diverse sources, ensuring it aligns with the research objectives and is comprehensive enough to yield meaningful insights.

Data Pre-processing: Raw data often requires cleaning, normalization, and transformation to enhance quality and usability. This step weeds out inconsistencies and prepares the dataset for analysis.

Subjecting Various Data Mining Techniques: Different data mining techniques such as clustering, classification, and association are applied to extract patterns, correlations, and structures from the pre-processed data[6].

Model Development Using Data Mining Techniques and Machine Learning Algorithms: Utilizing insights from data mining, machine learning algorithms like decision trees, neural networks, or support vector machines are employed to build predictive or descriptive models.

Performance Evaluation: Models are rigorously evaluated using metrics like accuracy, precision, recall, or F1-score to gauge their effectiveness in solving the research problem.

Comparative Analysis: Comparative studies assess the performance of different models or techniques to identify the most suitable approach based on predefined criteria.

Validation and Testing: Validation ensures the model's robustness and generalizability, often using techniques like cross-validation or hold-out validation. Testing involves applying the model to unseen data to simulate real-world scenarios.

Outcome and Result: The final stage reveals the outcomes and findings derived from the analysis. This includes actionable insights, discoveries, or predictions that contribute to the research domain or practical applications.

Each of these steps is integral to the iterative process of data analysis, ensuring that conclusions drawn are robust, reliable, and applicable in real-world contexts.

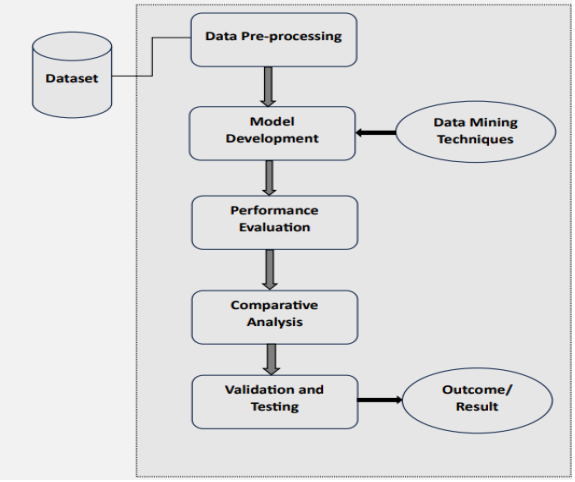

Figure 4.1. System Architecture design

V. PROPOSED WORK MODEL

DATASET COLLECTION

The Employment Scam Aegean Dataset (EMSCAD) sourced from Kaggle comprises 18,000 job descriptions, with approximately 800 identified as fraudulent postings. It includes job details and metadata. Our analysis utilizes this dataset to develop classification models aimed at identifying fake job postings[1]. The dataset features 18 columns, including the output column for classification purposes. The Table 5.1 represents the brief description of attributes in the dataset.

S. No.	Attribute	Description
1	Job ID	Unique Job ID
2	Title	The title of the Job
3	Location	Geographical location of the Job
4	Department	Corporate Department
5	Salary Range	Anticipated Salary Range
6	Company Profile	A brief description of the company
7	Description	A description of the job posting
8	Requirements	Minimum requirements for job opening
9	Benefits	List of benefits offered by employer
10	Telecommuting	Telecommuting position type
11	Has_company_logo	Presence of company logo
12	Has_Questions	Presence of screening questions
13	Employment type	Whether Full-Time. Part-Time or contract etc.
14	Required experience	Whether Executive, Entry Level or Intern etc.
15	Required education	Doctorate, Master's or Bachelor's etc.
16	Industry	Automotive, IT, Healthcare, Real estate etc.
17	Function	Engineering, Research, Marketing etc.
18	Target value (Fraudulent)	Target output of classification.

Table 5.1. Description of attributes of the Dataset

DATA PREPROCESSING

Initially, the dataset is analysed and removed columns that were not statistically significant. To handle missing data, we replaced empty values. Employed two techniques for data preprocessing: Under Sampling (chosen for this study) and Over Sampling to address data imbalance. After sampling and cleaning, the data is organized into a data frame. The dataset is categorized based on attributes like Country and Experience to examine data distribution. Generated two 'Word Clouds' for each category (Fake and Real outcomes) derived from the categorized dataset. The NLTK library helped identify common 'stopwords' for analysis. After categorization and analysis, the dataset is split into Train and Test sets using a fixed interval. Given that the input contained string features, we used 'Vectorization' to convert these features into vectors and then into matrices for experimentation. Throughout these steps, we utilized various Python modules and libraries for thorough data analysis, exploration, and preprocessing tasks. These procedures collectively ensured a structured approach to prepare the data for effective model experimentation and evaluation.

TRAIN-TEST SPLIT

After employing several Data Feature Extraction techniques and cleaning the dataset, the instances are split into two different sets. One of them is employed to train the model and another one to test the model. The feature values and target variable are considered respectively for both the training and testing sets.

DATA MINING TECHNIQUES EMPLOYED

Data Mining holds a variety range of techniques to extract patterns, forms, insights, and knowledge from extensive datasets. The methodologies briefly described below are employed in this study.

Sequential Pattern Mining: Discovering patterns or sequences where the order of occurrences matters. Techniques include sequence mining and temporal pattern mining.

Classification: Assigning predefined labels or categories to instances based on their features. Techniques include Decision Trees, Logistic Regression, Naïve Bayes model, and Support Vector Machine (SVM).

Anomaly Detection: Identifying unusual patterns or outliers in data that deviate from expected behaviour. Techniques include statistical methods, clustering-based anomaly detection, and supervised learning approaches.

Association Rule Mining: Discovering relationships and associations between variables in large datasets. The Apriori algorithm is widely used for mining association rules.

Dimensionality Reduction: Reducing the number of variables by selecting a smaller set, transforming, or combining them. Techniques include Principal Component Analysis (PCA) and feature selection techniques.

Regression Analysis: Predicting continuous numeric values based on the relationships between variables. Techniques include Linear Regression and Support Vector Regression.

Deep Learning: Utilizing neural networks with multiple layers to learn intricate patterns. Techniques include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models.

Clustering: Grouping similar instances together based on their characteristics without predefined categories. Techniques include K-Means Clustering and K-Nearest Neighbours.

Ensemble Methods: Enhancing predictive performance by combining multiple models. Techniques include bagging (Random Forest) and Stacking.

DETAIL OF ALGORITHMS DEPLOYED

- DECISION TREE

In data mining, a decision tree is a predictive modelling tool that maps out possible outcomes from a series of decisions. It organizes data into a tree-like structure where nodes represent decision points based on feature attributes, and branches represent the outcome of those decisions. Each decision node splits the data into subsets, guiding the tree towards a final prediction at the leaf nodes. Decision trees are popular due to their interpretability, allowing analysts to easily understand and visualize how decisions are made. They're used in classification and regression tasks, providing insights into complex datasets through a straightforward hierarchical framework.

- RANDOM FOREST

Random Forest is an ensemble learning method in data mining that constructs multiple decision trees during training. Each tree in the forest operates independently, making predictions based on random subsets of the features. The final prediction is determined by aggregating the predictions of all individual trees, often using voting for classification tasks or averaging for regression tasks. Random Forest mitigates overfitting and enhances accuracy by combining diverse models that capture different aspects of the data. It's robust against noise and outliers, making it a powerful tool for complex datasets where traditional models might struggle to generalize effectively.

- LOGISTIC REGRESSION

Logistic Regression is a statistical technique used in data mining and machine learning to predict the probability of a binary outcome (such as whether a customer will churn or not) based on one or more predictor variables. Despite its name, it's a classification algorithm rather than a regression algorithm.

In logistic regression, the dependent variable is binary or categorical, and the relationship between the independent variables (predictors) and the probability of the outcome is modelled using the logistic function. This function transforms the output into a range between 0 and 1, representing probabilities.

The model estimates the coefficients of the independent variables, which signify their impact on the probability of the outcome. These coefficients are derived through maximum likelihood estimation, fitting the model to the training data by minimizing the error between predicted and actual outcomes.

Logistic Regression is widely used for its simplicity, interpretability, and efficiency in handling large datasets. It's particularly useful when the relationship between predictors and outcomes is nonlinear or when dealing with categorical predictors. Applications include predicting customer churn, credit card fraud detection, and medical diagnostics, where understanding the likelihood of an event is crucial for decision-making and risk assessment.

- K – NEAREST NEIGHBOURS

K-Nearest Neighbours (K-NN) model is a simple yet effective supervised machine learning algorithm used for both classification and regression tasks. In K-NN, the prediction of a new data point is determined by the majority class (for classification) or the average (for regression) of its K nearest neighbours in the training data.

The algorithm operates on the principle of similarity, where similarity is defined using distance metrics such as Euclidean, Manhattan, or Minkowski distances. K-NN is non-parametric, meaning it doesn't make any assumptions about the underlying data distribution, making it versatile for various types of data.

One of the key parameters in K-NN is K, which determines the number of neighbours considered when making predictions. Choosing the right K is crucial as it affects the model's bias-variance trade-off; smaller

values of K tend to overfit the data, while larger values may underfit.

K-NN finds applications in recommendation systems, image recognition, and anomaly detection, where identifying similar patterns or objects based on their features is essential for decision-making.

- SUPPORT VECTOR MACHINE

Support Vector Machines (SVMs) are powerful supervised learning models used extensively in data mining for classification and regression tasks. SVMs work by finding the optimal hyperplane that best separates data points belonging to different classes. This separation is achieved by maximizing the margin, which is the distance between the hyperplane and the nearest data points (support vectors).

Mathematically, SVMs aim to solve an optimization problem that involves finding a hyperplane in a high-dimensional space that distinctly classifies the data points. The algorithm uses a kernel function to map the input data into a higher-dimensional space where a linear separation is possible, even if the original data are not linearly separable.

Key advantages of SVMs include their ability to handle high-dimensional data, effectiveness in cases where the number of dimensions exceeds the number of samples, and flexibility in choosing different kernel functions (e.g., linear, polynomial, radial basis function) to suit various data distributions.

However, SVMs can be sensitive to the choice of parameters such as the regularization parameter C and the kernel parameters, and they can be computationally intensive for large datasets. Despite these challenges, SVMs are widely used in applications such as image classification, text categorization, and bioinformatics due to their robust performance and theoretical foundations in machine learning.

Overall, SVMs are valued for their versatility, ability to handle complex decision boundaries, and robustness in various data mining tasks, contributing significantly to the advancement of predictive modeling in diverse domains.

- NAÏVE BAYE'S THEOREM

Naive Bayes is a probabilistic machine learning model widely used in data mining for classification tasks. It is based on Bayes' theorem, which calculates the probability of a hypothesis (class label) given the data. Despite its simplicity, Naive Bayes can be highly effective, especially in scenarios with limited training data.

Mathematically, Naive Bayes computes the posterior probability of each class given the input data by multiplying the likelihood of each feature given the class and the prior probability of the class, and then normalizing by the evidence across all classes.

Naive Bayes classifiers are particularly popular for text classification tasks, such as spam detection and sentiment analysis, where the independence assumption may hold reasonably well. They are also robust against irrelevant features and can handle categorical data efficiently.

While Naive Bayes may not capture complex relationships between features, its speed and simplicity make it a valuable baseline model and a go-to choice in many data mining applications, providing quick insights and reliable predictions in diverse domains.

- NEURAL NETWORK

Neural networks are foundational to data mining, leveraging their ability to model complex patterns and relationships in data through interconnected layers of artificial neurons. These networks are inspired by the human brain's neural structure, consisting of input, hidden, and output layers that process information sequentially.

In data mining, neural networks excel in tasks requiring pattern recognition, classification, regression, and even unsupervised learning. Each neuron receives inputs, applies a weighted sum and an activation function, and then passes the result to the next layer. This process enables neural networks to learn intricate mappings between inputs and outputs by adjusting the weights during training using optimization algorithms like gradient descent.

Deep neural networks (DNNs) extend this capability with multiple hidden layers, allowing them to capture increasingly abstract features from data. Convolutional neural networks (CNNs) specialize in processing grid-like data, such as images, by applying filters to extract spatial hierarchies of features. Recurrent neural networks (RNNs) are adept at handling sequential data, making them suitable for tasks involving time series or natural language processing.

Despite their power, neural networks require substantial computational resources for training, and their black-box nature can make interpretation challenging. Techniques like regularization, dropout, and batch normalization help mitigate overfitting and improve generalization.

Neural networks continue to drive innovations in data mining, revolutionizing fields such as computer vision, speech recognition, and autonomous systems. Their ability to learn from vast datasets and adapt to complex tasks underscores their importance in advancing artificial intelligence and data-driven decision-making across industries.

WORKFLOW OF THE MODEL

The algorithms are implemented and executed within a Python notebook, with the model being applied to the dataset after extracting features and identifying patterns through necessary pre-processing steps. Following the train-test split, dataset instances are used to classify authenticity of job posting using the models. Each model undergoes training with the training data and is prepared for prediction using the testing data. Real-life examples, in addition to the testing data, are also included in the experimentation phase. Accuracy scores, sensitivity, precision, F1-score, support, and recall metrics are calculated for each algorithm. Leveraging Python's capabilities, these scores are meticulously recorded, highlighting the most accurate model through comprehensive comparative analysis. Corresponding plots, matrices, and outputs are documented accordingly for thorough review and analysis.

VI. OUTCOME OF THE PROJECT

The outcome of this project involves creating an application aimed at assessing the accuracy and operational effectiveness of diverse Machine Learning algorithms employing various Data Mining methods. This empowers users to compare accuracy scores and conduct in-depth evaluations of each algorithm's performance. Additionally, it supports the prediction of model outputs using test data, specifically oriented towards classified outcomes. Through comprehensive comparative analysis, the goal is to identify the most effective and potentially accurate results. The application is specifically tailored for classifying job postings using the developed model. The comparative plot between algorithms is shown in Figure 6.1.

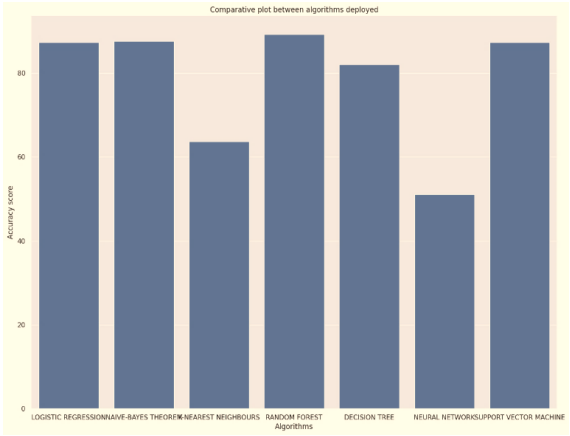

Figure 6.1. Comparison of accuracy scores of algorithms

This study helped us enhance our knowledge in Machine Learning Algorithms and the techniques of Data Mining and the way it organizes the synaptic signals for classification and progression of the data. The Table 6.1 represents the accuracy scores of the models in the study.

Model	Accuracy (%)
Logistic Regression	87.3
Naïve Baye's Theorem	87.59
K - Nearest Neighbors	63.64
Decision Tree	81.96
Neural Network	51.08
Support Vector Machine	87.3
Random Forest	89.1

Table 6.1. Output Scores of the algorithms

VII. APPLICATIONS

- Safeguard job seekers from scams by detecting and removing deceptive job postings that pose risks to personal or financial information.
- Automatically scan job postings on job boards to identify and prevent the publication of fraudulent postings.
- Enforce compliance with employment laws and regulations by filtering out postings that do not meet legal standards.

VIII. MERITS & DEMERITS

MERITS

- The model is capable to handle enormous number of datasets and attributes.
- Fast, efficient and accurate results can be produced.
- Cost-effective and time-efficient
- The model can further be developed using Deep Learning for enhanced applications.

DEMERITS

- The types of generation and source of fraud may not be established
- The severity of fraud may not be determined.

CONCLUSION

This paper encases the prediction of fake job posting using several proven techniques to establish patterns. has demonstrated substantial effectiveness in combatting the widespread issue of fraudulent job postings on online platforms. By employing machine learning algorithms such as Decision Trees, Random Forest Classifiers, and Naïve Bayes Models, a robust model capable of accurately identifying and flagging suspicious job listings based on textual feature analysis and semantic interpretation. Extensive evaluation and testing on real-world datasets confirmed the model's reliability, achieving around 90% accuracy. This underscores its robustness in distinguishing genuine opportunities from deceptive ones. The application of data mining techniques was pivotal in exploring and analysing the datasets, systematically interpreting results, defining algorithms for specific models, and training these models using meticulously gathered training data. As the Random Forest builds multiple decision trees, it ensures to compute the output from the majority of the trees that contribute to produce the efficient output accurately. The trees and the outputs are enhanced through proper training.

Hence, in regard to the subject, the fields of Data Mining and Machine Learning are the most emerging and beneficial in almost all the sectors and fields of studies. The models can be altered in the way they work for different kinds of datasets. By employing these kinds of techniques and computer predicted values, we can classify any data in efficient way with reduced cost.

BIBLIOGRAPHY

REFERENCES

[1] B. Alghamdi, F. Alharby, -An Intelligent Model for Online Recruitment Fraud Detection, Journal of Information Security, 2019, Vol 10, pp. 155 176.

[2] Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia- Tuan Nguyen, -Job Prediction: From Deep Neural Network Models to Applications, RIVF International

Conference on Computing and Communication Technologies (RIVF), 2020.

[3] Jiawei Zhang, Bowen Dong, Philip S. Yu, -FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network, IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[4] F. Murtagh, -Multilayer perceptron for classification and Regression, Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991.

[5] D. E. Walters, -Baye's Theorem and the Analysis of Binomial Random Variables, Biometrical J., vol. 30, no. 7, pp. 817 825, 1988

[6] P. Cunningham and S. J. Delany, -K -Nearest Neighbour Classifiers‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007

[7] 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