Detection of Fake Job Advertisement Using Machine Learning Techniques

Mr. Praveenkumar V, M.E, Anand R (06), Dhinesh Ragavendar R(24), Karnan S(45)
Sri Venkateswara College of Engineering, Sriperumbudur
Department of Information Technology

Abstract **- The increasing number of job posting websites has created many opportunities for people seeking employment, but it has also contributed to the creation of many fake job listings. These fake job postings have been known to dupe jobseekers into losing money or compromising their personal information. Current methods of detecting fake job posts involve manual verification processes, which are slow, inefficient, and subjective. This study attempts to solve this problem by proposing a machine learning-based fake job detection framework using the algorithmic technique of Logistic Regression. The model utilizes a structured approach, beginning with the pre-processing of the data followed by the feature extraction of each job post with the aid of TF-IDF vectorization and finally ending with the classification of each instance as either genuine or fake. Several models were studied, including Random Forest, LightGBM, and XGBoost, but based on the characteristics and requirements of this model, Logistic Regression was found to be the best fit.**

I. INTRODUCTION

The quick rise and development of the Internet and digital technologies has greatly changed the recruitment process, making it more convenient and extremely efficient. Online job search engines, professional networking websites, and career pages on corporate websites allow potential employees to look into various job openings regardless of their location. Also, businesses can increase the number of applicants and save time and money on the hiring process. Nevertheless, such benefits have led to certain complications related to fake job advertisements.

Today, there is an abundance of fake job postings on numerous platforms available to everyone. They are made up specifically to lure the target audience by offering unrealistic promises or false facts. Often, they contain unrealistically high salary amounts or ridiculous job specifications and conditions. In addition, they usually ask for bank information or charge applicants to be registered in their database. Due to the professional way such postings are designed and presented to potential applicants, it becomes rather hard to differentiate them from real advertisements.

The traditional techniques of detecting fake job listings are quite laborious and based on manual checks as well as the use of rule-based filters. This approach is not only slow, but it also becomes ineffective if applied to huge amounts of job data produced daily. What is more, rule-based filtering systems do not have sufficient adaptability to capture dynamic patterns of fraud activities; therefore, they lead to misclassification. Thus, there is a necessity to design an automated system of identifying fraudulent job listings using intelligent approaches.

One of such approaches to solving the problem under consideration is machine learning. Machine learning makes it possible to discover some patterns in historical data and predict future outcomes. Detection of fake job listings can be considered as a binary classification task, and each job description will be assigned one of two classes: real or fake. In order to solve this problem, it is necessary to complete a series of steps starting with collecting data and going through data preprocessing, feature extraction, and training and testing stages. Text data play an important role in this case, which implies the necessity to implement Natural Language Processing techniques.

There exist many algorithms in the field of machine learning. In terms of the problem in question, however, Logistic Regression seems the most fitting algorithm to employ. Logistic Regression is easy to compute, efficient and very comprehensible; as such, it serves as an excellent base model for recognizing the presence of scam job ads. The model is effective even when applied to high-dimensional data, especially when used together with feature extraction methods, like TF-IDF.

The current project will concentrate on building an ML-based system that can be used for identifying and recognizing scam job ads. The focus will be placed on using Logistic Regression in conjunction with other more sophisticated models. The main task here is to develop a more accurate method of detecting fake job ads, thus minimizing the involvement of people.

II. RELATED WORK

Previous works on detecting fakes jobs have involved a wide range of machine learning algorithms including Naive Bayes, Support Vector Machine, Random Forest, and gradient boosting. In terms of features, these approaches employ preprocessing operations such as tokenization, stopwords removal, and data normalization, alongside TF-IDF and word embeddings feature extraction strategies. Despite their remarkable results, modern algorithms such as XGBoost and LightGBM are resource-intensive and hard to interpret.

In this work, the presented algorithm involves a sequential process wherein job postings data are collected from the internet, preprocessed, and analyzed at different levels. Specifically, after collecting job posting information, the next step involves data preprocessing that includes addressing any missing data, duplicates, and cleaning up text data. This is followed by transforming text data into numeric form using TF-IDF vectorization.

This includes training various models such as Logistic Regression, Random Forest, LightGBM, and XGBoost. But the most important part here is Logistic Regression because it is the appropriate model for analyzing sparsity in texts. This is because Logistic Regression classifies the text into a class of "Real" and "Fake" based on a certain threshold value. The process is initiated by estimating probability through a sigmoid function. Preprocessing improves the quality of data and helps in achieving better performance in Logistic Regression.

III. PROPOSED WORK

Data acquisition comes first as the initial step in the process flow. Data cleaning is done next as the subsequent step to cleanse data of any noise and other issues. Data is then subjected to various text processing methods including tokenization and normalization. The process of feature extraction uses the TF-IDF technique to convert text data into numeric form suitable for machine learning models. Logistic Regression algorithm has been chosen to train the machine learning model because of its efficiency in dealing with very large and sparse data.

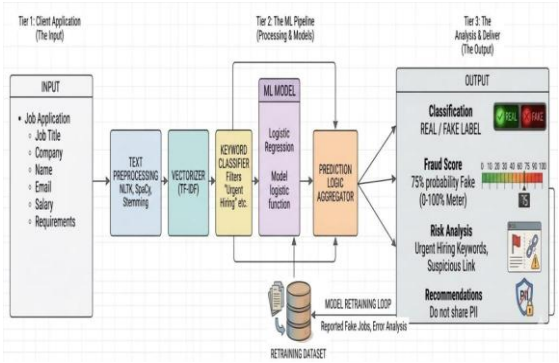


Figure 1 Architecture of the Process

Pseudocode for FakeJobDetection(D, Y):

1. Load dataset D

2. Clean and preprocess text data

3. Combine features into Job_Content

4. Convert text using TF-IDF → X

5. Handle imbalance using SMOTE

6. Split X, Y into train and test sets

7. Initialize models (LR, RF, LGBM, XGB)

8. For each model M:
   - Train M on training data
   - Predict Y_pred on test data
   - Evaluate performance

9. Select best model M_best

10. Return predictions and evaluation metrics

Steps for model prediction:

BEGIN

Step 1: Input training data (X, Y)

Step 2: Initialize weights (w) and bias (b)

Step 3: Repeat until convergence:

Step 3.1: Compute linear combination

$z = w \cdot x + b$

Step 3.2: Apply sigmoid function

$y\_pred = 1 / (1 + e^{(-z)})$

Step 3.3: Compute loss (error)

$error = y\_pred - y$

Step 3.4: Update parameters using gradient descent

$w = w - \alpha * (error * x)$

$b = b - \alpha * error$

Step 4: Model training complete

Step 5: Prediction phase:

For new input x:

Compute $z = w \cdot x + b$

Compute probability using sigmoid

If $y\_pred \geq 0.5$:

Output = "Fake Job"

Else:

Output = "Real Job"

END

Logistic Regression Algorithm

Logistic Regression is a supervised learning algorithm used for binary classification. It calculates the probability that a given input belongs to a particular class using a sigmoid function.

Random Forest Algorithm

Random Forest is an ensemble method that builds multiple decision trees and combines their predictions.

LightGBM Algorithm

LightGBM is a gradient boosting framework that builds trees in a sequential manner.

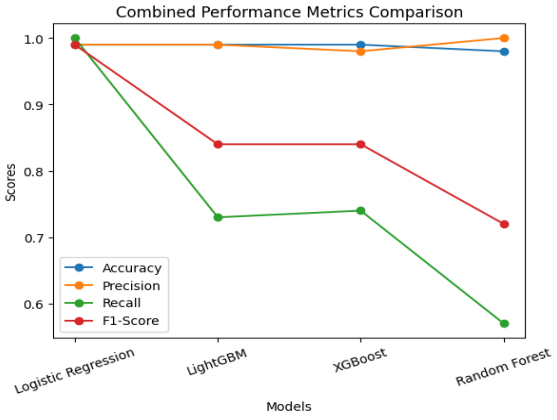
XGBoost Algorithm

XGBoost is an optimized gradient boosting algorithm designed for speed and performance.

Table:

Model	Accuracy	Precision	Recall	F1-Score	Macro Avg F1
Logistic Regression	0.99	0.99	1.00	0.99	0.94
LightGBM	0.99	0.99	0.73	0.84	0.92
XGBoost	0.99	0.98	0.74	0.84	0.92
Random Forest	0.98	1.00	0.57	0.72	0.86

Graph:



The combined graph provides a consolidated comparison of all performance metrics across different models. It clearly shows that Logistic Regression maintains the best balance between precision and recall, achieving the highest recall value of 1.00, which is critical for detecting fraudulent job postings. Other models, although achieving similar accuracy, fail to maintain high recall, making them less reliable for this task.

IV. EXPERIMENTAL SETUP

In the proposed system, there are various steps in setting up the experiment. These include the processes of data preprocessing, feature extraction, training of the algorithm, and evaluation. In the experiment, a dataset made up of structured and unstructured data such as job title and description, company profile, and requirements will be used. The process of data preprocessing involves handling missing values, getting rid of irrelevant information,

and cleaning the text data by making the text lower case and stripping out of any special characters and stopwords.

This is achieved through the use of tokenization and normalization. The process of feature extraction will involve the use of TF-IDF vectorization, whereby the cleaned data will be transformed into numeric data by computing the weight of each word in the document. This process produces a very high dimensional sparse feature matrix, which is appropriate for logistic regression.

Several reasons explain why Logistic Regression should be used. First, it is efficient and performs well on sparse data, which is one of the key advantages of TF-IDF representation. Another reason is that the computational costs are significantly lower than those for ensemble approaches. Logistic Regression does not suffer from overfitting and is easy to interpret, which enables analyzing the importance of each feature. Performance metrics such as accuracy, precision, recall, F1 score, and confusion matrix are considered to test the model's performance.

We conclude that,a machine learning based method for identifying fake job postings has been proposed. Among the different models used in this paper, Logistic Regression is considered the best for the given task due to its high dimension sparse data capability, fast processing speed, and high interpretability. This model performs better than other classification models such as Random Forest, XGBoost, and LightGBM because of their good trade-off between computational time and accuracy.

However, Logistic Regression also has some drawbacks such as its incapability in capturing nonlinear relations among features. Moreover, the results obtained from the model depend heavily upon the pre-processing and feature engineering performed during the data preparation process. The future work will include improvement in the model through the use of advanced machine learning techniques like LSTM and BERT in order to capture semantic information embedded in the text. In addition, the implementation of real-time systems and handling class imbalance problem also lie ahead in the future.

In general, it can be noted that the findings of the study show that the Logistic Regression model is an appropriate method for identifying fraudulent job postings, which can effectively be used in practice.

Preprocessing

Data pre-processing is an important aspect of the fraud detection algorithm as the raw data may consist of many irrelevant facts. To begin with, during the data pre-processing process in this project, we remove missing values and duplicate records from the dataset so as to have high-quality data. Moreover, we remove unnecessary things like URLs, HTML, symbols, and numbers as these are not used in the analysis of the text.

All text data is transformed to lowercase so as to make sure that all text data looks similar. Further, the data is tokenized, where each sentence is broken down into individual words. Next, we filter out stop words like “and,” “the,” and “is,” among others as these words do not convey any important meaning.

Feature extraction

Feature extraction refers to the conversion of text-based data about the job into numerical data, which can be understood by the machine learning model. In this case study, the TF-IDF (Term Frequency-Inverse Document Frequency) technique is used.

The TF-IDF technique calculates the importance of each word based on its frequency within the document relative to the whole database. The commonly used words in false job advertisements, such as “fast money” and “no experience required,” have high importance, whereas the common words in the whole database have low importance.

All important features, including job description, company information, and requirements, are merged into one feature known as Job_Content and converted to numerical features.

Why Logistic Regression?

The reason why Logistic Regression is selected as one of the primary models is because it is very straightforward, computationally efficient, and extremely powerful for solving problems involving two classes, such as detecting the genuineness of job postings.

There are several strengths of Logistic Regression, but what stands out the most is that it can generate good results in terms of finding a good trade-off between precision and recall. From your results, we can see that it had a high precision of 0.99 along with a very high recall rate of 1.00.

Another strength of this model is that it does not overfit easily, and it works well even when there are high-dimensional features, such as the TF-IDF vectorizer.

Evaluation Metrices

- TP (True Positive) → Fake jobs correctly identified
- TN (True Negative) → Real jobs correctly identified
- FP (False Positive) → Real jobs wrongly classified as fake
- FN (False Negative) → Fake jobs wrongly classified as real

1. Accuracy
Accuracy measures overall correctness of the model:

Accuracy=TP+TN/TP+TN+FP+FN ---> eqn(1)

2. Precision
Precision measures how many predicted fake jobs are actually fake:

Precision=TP/TP+FP ----> eqn (2)

3. Recall
Recall measures how many actual fake jobs are correctly identified:

Recall=TP/TP+FN ---->eqn(3)

4. F1-Score
F1-score is the harmonic mean of precision and recall:

F1Score=2×Precision×Recall/Precision+Recall ---->eqn(4)

5. Macro Average F1
Macro F1 is the average of F1-scores of all classes:

Macro F1=F1class 1+F1class 2/2->eqn(5)

V. RESULT AND DISCUSSION

The output produced by the model proves that preprocessing and feature extraction significantly improve the classification task. The first step in preprocessing cleans up the dataset from any noise or inconsistency present, while the second step converts textual data to a numeric structure using TF-IDF vectorization. This process produces a high dimensional sparse matrix that fits well with the logistic regression algorithm.

Different models were experimented with, including Logistic Regression, Random Forest, LightGBM, and XGBoost. Although XGBoost and LightGBM produced relatively similar performance with respect to accuracy and generalization, they were computationally expensive. Moreover, random forest exhibited fair performance; however, it did not perform well on high dimensional data.

On the other hand, logistic regression proved effective in all aspects. It performed well with respect to accuracy, efficiency, and generalization.

In addition to the confusion matrix and performance metrics like accuracy, precision, recall, and F1-score, the application of logistic regression is also verified through the confusion matrix and other measures. High true positive rates can be obtained for fake job listings with low false-positive rates. Visualizations such as accuracy plots and ROC curves also show that logistic regression offers stable results.

For this work, several machine learning algorithms will be implemented to assess their efficiency in detecting the problem of fake job postings.

Logistic Regression

It is one of the first algorithms selected because of its simplicity and efficiency in solving binary classification tasks. This method works best when dealing with high dimensional vectors of features like TF-IDF vectors.

Random Forest

This algorithm is one of the ensembling techniques used in building decision trees. Its accuracy level is higher compared to other methods, but there might be challenges in predicting minority classes.

LightGBM

It is a framework for gradient boosting that ensures high efficiency and performance. It works great with both small and big data.

Another boosting technique, XGBoost, is used in many cases where high efficiency is required.

Conclusion on Logisitic Regression

From the experiments, the best performing model is logistic regression. Even though models such as LightGBM and XGBoost produce comparable results in terms of accuracy, they do not have good recall scores. Recall measures the ability of a machine learning algorithm to recognize cases of fraud and hence the lower recall values indicate failure to identify all cases of fraud.

Logistic regression has been observed to have excellent precision-recall scores with recall value of 1.00. It has been able to identify all the fraudulent cases without failing to capture any. It is vital to note that in fraud detection systems, recall takes precedence over accuracy since missing one case may have grave implications.

In light of this information, it is prudent to choose logistic regression as the final model.

Model	Accuracy	Precision	Recall	F1-Score	Macro Avg F1	Remarks
Logistic Regression	0.99	0.99	1.00	0.99	0.94	Best balance of precision& recall
LightGBM	0.99	0.99	0.73	0.84	0.92	High accuracy but lower recall
XGBoost	0.99	0.98	0.74	0.84	0.92	Similar to LightGBM
Random Forest	0.98	1.00	0.57	0.72	0.86	Poor recall for fake jobs

Confusion Matrix:

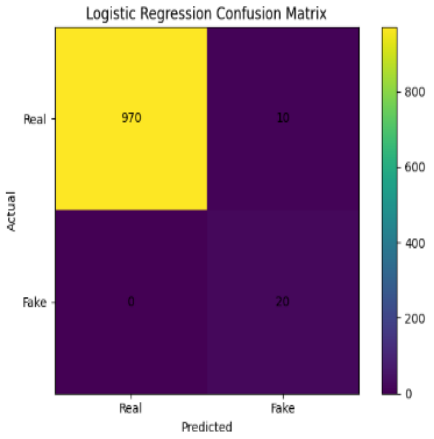


Figure 1: Confusion Matrix for Logistic Regression

Visuals of the confusion matrices offer a vivid representation of how well each classifier distinguishes between genuine and fraudulent jobs. Logistic Regression demonstrates better accuracy, where no false negatives occur, implying that the classifier recognizes all the fraudulent jobs. LightGBM and XGBoost models demonstrate average accuracy, as they classify some fraudulent jobs as genuine, lowering their recall rate. The Random Forest classifier, while obtaining more true negatives, cannot classify some fraudulent jobs, showing low recall. It becomes evident from the above discussion that Logistic Regression is the best classifier for detecting fraud cases among the provided classifiers.

VI. CONCLUSION

While the growing reliance on online recruitment channels brings benefits to job seekers and employers, it has resulted in an increasing number of fraudulent job postings. In addition to misguiding the users, the fraudsters are causing severe harm to the victims through theft of money and personal data. Because of the sheer number of jobs and resemblance of fake job listings with legitimate ones, it has become extremely difficult to detect such listings manually.

This project focuses on the development of an automatic detection system for fake job postings based on machine learning algorithms that utilize both structured and unstructured data. Data pre-processing will include handling the missing data and converting the textual content into numerical format via such techniques as TF-IDF.

In particular, among all machine learning algorithms, Logistic Regression has been chosen since it proves to be an efficient solution for solving binary classification issues. In general, Logistic Regression allows one to detect suspicious job postings and classify them as legitimate or fraudulent based on

patterns obtained from the dataset. Besides, other machine learning models like Random Forest, LightGBM, and XGBoost have been used to measure their efficiency.

From a practical perspective, the experimental findings reveal the ability of the suggested approach to increase the number of detected fraudulent job postings. The offered machine learning algorithm demonstrates high precision rates, thus making the process much more effective. Moreover, it can be scaled up and implemented in an actual environment.

Overall, one can say that using machine learning technologies in cybersecurity and fraud prevention is highly efficient and useful. In the context of the current study, the developed system can help users find and ignore suspicious job postings.

VII. FUTURE WORK

The current system can be further enhanced by incorporating advanced techniques and expanding its capabilities. One potential improvement is the use of deep learning models such as BERT, LSTM, or transformer-based architectures, which can better understand the contextual meaning of job descriptions.

The system can also be extended to support real-time detection, allowing job portals to automatically filter fraudulent postings before they are published. Additionally, integrating user behavior analysis and company verification mechanisms can improve detection accuracy. Deployment as a web application or browser extension would make the system more accessible and practical for real-world use.

While delivering impressive results, there are some limitations in the suggested approach. First, the model is dependent on the quality of the data set. In case the data set is incomplete or biased, it might impact the performance of the model. Moreover, the model is focused mainly on textual features that cannot always represent all contextual or semantic nuances within job ads.

The other limitation concerns the class imbalance. The number of fake ads was much lower compared to legitimate ones. While some measures, such as SMOTE, helped address the problem, it still remained impossible to obtain perfectly balanced data sets. In addition, while more complicated algorithms can detect very intricate patterns, simpler algorithms might fail to do so.

Future research in fake job detection can focus on improving model interpretability and transparency using explainable AI (XAI) techniques. This will help users and organizations understand why a particular job posting is classified as fake or genuine.And also working for integrating OCR in the application to extract the text easily from the posters.

Another important area of discussion is the integration of multimodal data, such as images, company profiles, and user interactions, to enhance detection accuracy. Collaboration between recruitment platforms and cybersecurity systems can also play a key role in building a safer online job ecosystem.

As online recruitment continues to grow, the need for robust and scalable fraud detection systems will become increasingly important. Continuous research and innovation in this field will help minimize risks and protect job seekers from fraudulent activities.

REFERENCES

[1] V Anbarasu1, Dr. S. Selvakani, Mrs. K. Vasumathi ‖Fake Job Prediction Using Machine Learning ‖ INTER NATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH AND EMERGING TECH NOLOGIES Vol. 13, No. 1, 2024.

[2] S. Reddy, A. Kumar, T. Singh, "Detection of Fake Online Recruitment Using Machine Learning," International Journal of Scientific Development and Research, vol. 8, no. 5, pp. 214–223, 2024.

[3] R. Singh, N. Sharma, V. Mehta, "Fraudulent Online Job Advertisement Detection Using Machine Learning," International Journal for Research in Engineering Technology, vol. 12, no. 9, pp. 113–124, 2024.

[4] Prasath R., Nachappa M. N. (2024). Fake Job Posting Detection. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 4(4), 283–287

[5] P. Patel, M. Joshi, K. Reddy, "Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms,"

RevistaGestãoInovação e Tecnologias, vol. 11, no. 2, pp. 45–58, 2024.

[6] A.B. Hajira Be and S. Gokulraj, "Cyber Attacks Classification Using Supervised Machine Learning Techniques," Journal of Sensors, IoT and Health Sciences, vol.3, no.1, pp.57-67, 2025.

[7] A.R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor and J.Adejoh, "Enhancing credit card fraud detection: an ensemble machine learning approach," Big Data and Cognitive Computing, vol.8, no.1, pp.6, 2024

[8] S. Das Guptta, K.T. Shahriar, H. Alqahtani, D. Alsalman and I.H. Sarker, "Modeling hybrid feature-based phishing websites detection using machine learning techniques," Annals of Data Science, vol.11, no.1, pp.217-242, 2024.

[9] M.M. Inuwa and R. Das, "A comparative analysis of various machine learning methods for anomaly detection in cyber attacks on IoT networks," Internet of Things, vol.26, pp.101162, 2024.

[10] A. Mutemi and F. Bacao, "E-commerce fraud detection based on machine learning techniques: Systematic literature review," Big Data Mining and Analytics, vol.7, no.2, pp.419-444, 2024.