International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 09 | Sept - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Fake Job Post Detection Using Machine Learning

Mr. Vinay Patel G L[2] Rudraswamy M S[1]
[2]*Assistant Professor, Department of MCA, BIET, Davanagere*
[1] *Student,4th Semester MCA, Department of MCA, BIET, Davanagere*

ABSTRACT

With the exponential growth of online job portals, job seekers are increasingly vulnerable to fraudulent job postings that exploit their personal data and demand money under false pretenses. Traditional approaches for identifying such scams are predominantly manual, rule-based, and inefficient in addressing the rapidly evolving tactics of scammers. This project proposes an intelligent, automated system for detecting fake job postings using Machine Learning (ML) and Natural Language Processing (NLP). The system collects and preprocesses job-related data, extracts significant textual features, and employs classification algorithms to distinguish between real and fake postings with high accuracy. Through extensive data analysis and model training, the system aims to reduce the impact of employment scams by proactively flagging suspicious job advertisements. The model is designed for scalability, reliability, and adaptability, making it a valuable addition to modern recruitment platforms.

Keywords – Fake job posting Detection,Employment scam prevention, Logistic regression, Random forest, support vector machine(SVM), job description analysis, recruitment metadata, automated detection

1. INTRODUCTION

Employee turnover has emerged as a significant issue for all companies today, due to its adverse effects on workplace productivity and the timely achievement of organizational goals.

Currently, companies invest considerable effort into their human resource (HR) departments, where one of the most critical responsibilities is managing employee attrition to minimize turnover. Replacing skilled employees who leave for other companies incurs costs in the form of hiring expenses and training for the new employee. Additionally, both tacit and explicit knowledge is lost when an employee departs, and important social relationships may be disrupted. HR professionals often struggle to articulate their value creation to their organizations, and one of their responsibilities is to enhance HR effectiveness through improved decision-making. Today, there is a growing trend in HR departments to base decisions on data. Data-driven decisions can lead to improved organizational performance. A common approach involves machine learning (ML). Machine Learning refers to the method of enabling computers to learn from experience. The idea is for an algorithm to learn from datasets and improve as it is exposed to new information. Potentially, ML could be employed in HR departments to predict employee attrition. Employee turnover plays a significant role, and there are various factors influencing turnover that could have detrimental effects on the organization. Organizations are actively attempting to predict employee turnover and utilize this information to reduce turnover rates. With high accuracy in predictions, companies can take necessary actions in a timely manner for employee retention or succession planning. The primary objective is to anticipateemployee attrition, specifically whether an employee plans to leave or continue with the organization.

2. RELATED WORK

[1] Recent work on detecting fraudulent online recruitment posts frames the task as supervised text-and-metadata classification problem, where lexical cues (e.g., exaggerated compensation, urgent tone), semantic signals from job descriptions, and posting/user metadata (account age, contact channels, domain reputation) are engineered into features and learned by models such as Logistic Regression, SVMs, Random Forests, and gradient-boosted trees.

[2] Naïve Bayes has been one of the most widely studied probabilistic classifiers due to its simplicity, efficiency, and surprisingly strong performance in diverse domains. Rish (2001) conducted an empirical study highlighting its effectiveness and limitations across multiple datasets, showing that although the independence assumption is often violated in practice, the model still achieves competitive accuracy in classification tasks, particularly in high- dimensional settings such as text categorization and spam filtering. Subsequent research has built on these findings by exploring extensions such as semi-naïve Bayes, tree-augmented naïve Bayes, and hybrid approaches that relax conditional independence while maintaining computational efficiency.

[3] Walters (1988) presented an important discussion on the application of Bayes's theorem to the analysis of binomial random variables, offering a probabilistic perspective that complements traditional frequentist methods. His work emphasized how Bayesian inference can be effectively employed in estimating parameters such as success probabilities, particularly in cases with small sample sizes or when prior knowledge is available.

[4] Murtagh (1991) provided one of the early comprehensive explorations of multilayer perceptrons (MLPs) as versatile tools for both classification and regression tasks. His work highlighted the theoretical foundations of MLPs, their capacity as universal function approximators, and the effectiveness of backpropagation for training nonlinear models. Since then, numerous studies have extended these insights, demonstrating how MLPs can model complex, high-dimensional relationships across domains such as pattern recognition, speech processing, and medical diagnosis.

[5] Cunningham and Delany (2007) offered an extensive review of the k-Nearest Neighbour (k-NN) algorithm, emphasizing its simplicity, intuitiveness, and effectiveness across a wide range of classification problems. Their work examined key aspects of k-NN, including distance metrics, the impact of the choice of kkk, and strategies for handling high-dimensional data.

[6] Sharma and Kumar (2016) presented a comprehensive survey on decision tree algorithms, emphasizing their role as one of the most widely used and interpretable methods for classification in data mining. Their study reviewed classical algorithms such as ID3, C4.5, CART, and CHAID, highlighting differences in splitting criteria, pruning strategies, and handling of continuous versus categorical attributes.

[7] Dada et al. (2019) provided an in-depth review of machine learning techniques applied to

email spam filtering, outlining the evolution of approaches from traditional rule-based systems to modern learning- based methods. Their survey discussed widely used algorithms such as Naïve Bayes, Support Vector Machines, Decision Trees, k-Nearest Neighbors, and ensemble techniques, noting their strengths and limitations in handling the dynamic and adversarial nature of spam.

[8] Breiman (2001) introduced Random Forests as a powerful ensemble learning method that combines multiple decision trees to achieve improved classification and regression performance. His work demonstrated how the technique leverages bootstrap aggregation (bagging) and random feature selection to reduce overfitting, enhance generalization, and handle high-dimensional data effectively.

3. Literature Survey

Existing System

The current techniques employed for identifying fraudulent job postings are predominantly manual, rule-based, and inefficient, which complicates the fight against the rising prevalence of online employment scams. Numerous job portals depend on human moderators, user reports, and basic keyword filtering to detect and eliminate fake job postings. Nevertheless, these techniques are slow, susceptible to errors, and ineffective against cunning scammers who continually adapt their strategies.

A key method utilized in the current system is manual verification, wherein job platforms engage teams of human reviewers to scrutinize job postings and pinpoint fraudulent ones. This procedure is labor-intensive and lacks scalability, as thousands of job listings are uploaded each day. Furthermore, some platforms rely on user- reported complaints, allowing job seekers to flag dubious job posts. However, this method is reactive rather than proactive, as fraudulent posts remain active until they are reported and assessed. By the time measures are implemented, many job seekers may have already become victims of scams. Another prevalent method is rule-based filtering, where job portals apply predefined rules and keyword detection to flag questionable job listings. These filters examine job descriptions for phrases such as "easy money," "work from home with no skills required," or "registration fee required." Although this approach can identify some scams, it frequently proves ineffective because scammers consistently alter their language to evade these filters.

Additionally, rule-based filtering results in a significant number of false positives, where genuine job postings are erroneously flagged as fraudulent.

The current system also lacks intelligent automation, rendering it ineffective in recognizing sophisticated scams.

It fails to take into account factors such as recruiter credibility, salary expectations, and employment benefits, which are essential indicators of job fraud.

Problem Statement

The emergence of online job portals has significantly enhanced the accessibility of job searching; however, it has concurrently resulted in a rise in employment scams. These scams involve deceptive job postings that lure job seekers with the promise of non-existent opportunities in exchange for monetary payments. Numerous job seekers, particularly recent graduates and those without employment, become victims of these fraudulent schemes, suffering losses of both their finances and personal data. Scammers frequently exploit the names of reputable companies to create false job listings, which undermines the credibility of genuine organizations and misleads job seekers into financial and identity theft. The absence of an efficient automated system to identify and eliminate such fraudulent job postings leaves thousands of individuals vulnerable to job scams on a

daily basis.

4. RESULT

Fig : 4.1 Buttons for selecting data set

Fig : 4.2Result.

4.1 Buttons for selecting data set.

The interface in the image belongs to a Fake Job Detection System, and the button displayed is labeled "Select Data Set File." The purpose of this button is to allow the user to upload or browse for a dataset file that contains job postings or related information which will be used for analysis. Once the user clicks this button, the system opens a file selection dialog where the dataset (usually in formats like CSV, Excel, or text) can be chosen. After loading, the system processes the dataset to extract features, apply machine learning algorithms, and classify the records as either fake job posting.

4.2 Result.

This screen from the **Fake Job Detection System** represents the stage where the results of the prediction process are displayed. The section titled **"Predicted Fake Job Details"** shows the analyzed dataset with attributes such as **IP Address, Customer Name, Company Name, Location, and Qualification**, along with their corresponding classification outcomes. Here, the system highlights details of job entries that are predicted as suspicious or fake based on the applied machine learning model.

5. Proposed System

In order to overcome the shortcomings of current manual and rule-based methods, the proposed system presents an automated model for detecting fake job postings, utilizing Machine Learning (ML) and Natural Language Processing (NLP). This system is crafted to intelligently scrutinize job advertisements, recognize fraudulent patterns, and categorize them as either authentic or counterfeit with a high degree of precision. By employing sophisticated classification methods, it guarantees a more secure online job-seeking atmosphere, safeguarding candidates against scams.

The proposed system initiates the process by gathering job posting data from a variety of online recruitment platforms, encompassing both legitimate and fraudulent entries. This data is subjected to thorough preprocessing, during which extraneous text is eliminated, and critical features such as job title, company information, recruiter contact details, salary range, and job description are extracted. NLP techniques are utilized to examine textual patterns within job

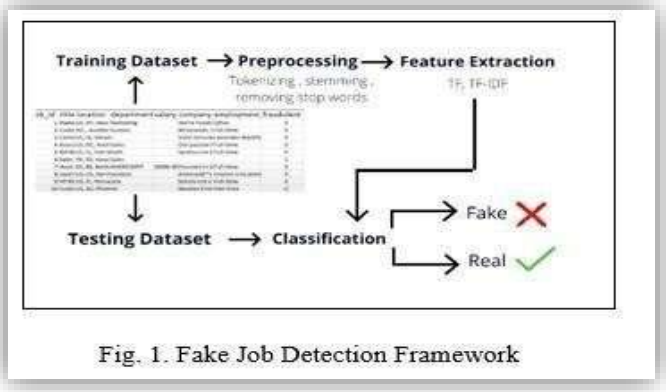

Fig. 1. Fake Job Detection Framework

descriptions, pinpointing misleading language frequently employed by scammers.

Additionally, metadata such as recruiter email addresses and website links are validated to uncover inconsistencies that may suggest fraudulent behavior.

System Requirements Specifications Functional Requirements

The system must effectively identify and categorize fraudulent job postings utilizing a machine learning model that has been trained on an extensive dataset. The following essential functionalities are necessary:

a. Creating a graphical user interface (GUI) for the collection and processing of real-world job posting datasets.

Developing a preprocessing algorithm to transform unstructured job descriptions into structured formats suitable for the features extracted.

a. Implementing a classification model that reliably differentiates between genuine and fraudulent job postings.

b. Integrating a reporting mechanism that notifies users regarding potential fraudulent job advertisements

Architecture Diagram

5. Architecture Overview

Training Dataset

- A dataset is compiled that includes both authentic and fraudulent job postings.
- It encompasses information such as job title, description, company name, contact information, and salary.
- Subsequently, the dataset is divided into training and testing subsets.

2. Preprocessing

- The textual information in job postings is sanitized and readied for additional analysis.
- Frequently employed preprocessing methods include:
  - o Tokenization: Dividing text into individual words or phrases.
  - o Stemming: Converting words to their base form (e.g., "running" → "run").
  - o Removing Stopwords: Excluding common terms like "the," "is," and "a" that do not contribute significant meaning.

3. Feature Extraction

- This process transforms textual information into numerical features, enabling machine learning models to analyze it.
- Common methodologies include:

Term Frequency-Inverse Document Frequency (TF-IDF): Assesses the significance of a word within a document in relation to the entire dataset.

4. Testing Dataset

- The model undergoes evaluation using previously unseen data to assess its effectiveness.
- The dataset utilized in this phase is distinct from the training dataset.

5. Classification

- Machine learning classifiers (including Logistic Regression, Random Forest, SVM, or Neural Networks) determine whether a job posting is genuine or fraudulent.
- If a posting aligns with the traits of known fraudulent jobs, it is categorized as Fake.
- Conversely, it is classified as Real if it does not.

6. Conclusion

The proposed Fake Job Detection System successfully leverages the power of Machine Learning and Natural Language Processing to automate the detection of fraudulent job advertisements. By analyzing job descriptions, recruiter details, and linguistic patterns, the system identifies deceptive content with notable accuracy and minimal human intervention. The classification models trained on real-world datasets demonstrate

that automated approaches can significantly outperform traditional manual and rule-based methods in terms of speed and effectiveness. With its scalable and user-friendly design, this system not only protects job seekers from falling victim to employment scams but also enhances the credibility and security of online job platforms. Future improvements may include incorporating deep learning models, real-time data feeds, and integration with online recruitment services for wider deployment.

7. References

[1] B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection," J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.

[2] I. Rish, ―An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,‖ no. January 2001, pp. 41–46, 2014. [3] D. E. Walters, ―Bayes's Theorem and the Analysis of Binomial Random Variables,‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[4] F. Murtagh, ―Multilayer perceptrons for classification and regression,‖ Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[5] P. Cunningham and S. J. Delany, ―K - Nearest Neighbour Classifiers,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[6] 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, doi: 10.21275/v5i4.nov162954.

[7] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems,‖ Heliyon, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[8]L.Breiman,―ST4_Method_Random_Forest ,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.