JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Fake Job Classifier Using Machine Learning

A Phase 1 Report on Fake job Classifier Using Machine Learning

[1]Ms.THEEKSHA K, [2]Dr. Shashidhar Kini K

[1]Student, [2]Professor & Head
[1]Department of Master of Computer Applications,
[1]Srinivas Institute of Technology, Valchil Mangaluru, Karnataka, India.

Abstract : *Abstract-Online job portals have become primary platforms for employment opportunities. However, with their popularity comes an increase in fraudulent job postings designed to exploit job seekers. These scams can lead to financial losses, identity theft, and psychological stress. This paper presents a machine learning-based system for classifying fake job postings using Natural Language Processing (NLP) techniques. By extracting linguistic and metadata-based features from job descriptions, our proposed model leverages classifiers such as Logistic Regression, Random Forest, and XGBoost. We used the Employment Scam Aegean Dataset (EMSCAD) for training and evaluation. Experimental results demonstrate that the XGBoost model provides the best performance with high precision and recall, showcasing its effectiveness in detecting fraudulent listings.*

.

IndexTerms - Fake job classification, machine learning, NLP, XGBoost, EMSCAD, text mining

I. INTRODUCTION

In recent years, the internet has revolutionized the recruitment industry, making job search platforms and online career portals more accessible than ever before. However, this digital convenience has also opened doors for malicious entities to exploit unsuspecting job seekers through fake job postings. These fraudulent listings often involve deceptive offers, phishing schemes, and requests for sensitive personal or financial information. Such scams not only waste time and resources but can also lead to severe psychological and financial harm to applicants.

Manual moderation and rule-based filters on job platforms are often inadequate to detect these scams due to their evolving nature and the subtlety with which they mimic legitimate postings. Therefore, there is a growing need for intelligent, automated systems that can accurately identify and flag fake job listings. Machine learning, combined with natural language processing (NLP), offers a powerful approach to analyzing large volumes of job data and learning patterns that distinguish genuine posts from fraudulent ones.

This research aims to develop a fake job classification model using supervised machine learning techniques. By preprocessing and analyzing textual and categorical data from real-world job postings, the model learns to predict the authenticity of a job listing. The ultimate goal is to build a tool that can be integrated into job portals to enhance user safety, improve trust, and minimize the risk of online employment fraud.

II. EASE OF USE

The fake job classification system is developed to be easily accessible and user-friendly, enabling seamless integration into job portals, recruitment platforms, or standalone web applications. Users can input job details such as title, description, company profile, and employment type through a simple interface. Upon submission, the model processes the data and returns a clear classification— Fake or Real—along with a confidence score that indicates the reliability of the prediction. This immediate feedback allows job seekers and administrators to make informed decisions before proceeding with an application or listing.
Designed for scalability and adaptability, the system can process both individual and bulk job postings efficiently. It supports periodic retraining to incorporate newly emerging patterns of job fraud, ensuring continued accuracy over time. With minimal technical expertise required for use, and the potential for multilingual support, the system enhances safety for a wide range of users across different regions and platforms. Its lightweight design also makes it suitable for deployment on web and mobile applications alike.

Prepare Your Paper Before Styling

Before formatting the research paper, significant effort was devoted to preparing a clean, structured, and high-quality dataset suitable for machine learning tasks. The initial step involved acquiring the dataset from Kaggle, which included over 17,000 job listings labeled as either genuine or fake. Key attributes such as job title, company profile, job description, requirements, employment type, and benefits were selected for analysis. These fields, especially the textual data, required extensive preprocessing to ensure consistency and clarity.

Preprocessing included removing missing values, HTML tags, special characters, and irrelevant symbols. Natural Language Processing (NLP) techniques such as lowercasing, stop word removal, stemming, and TF-IDF vectorization were applied to extract meaningful features from the job descriptions and company profiles. Categorical features like employment type and required experience were encoded using label and one-hot encoding techniques. The dataset was then split into training and testing sets to evaluate different classification models.

After the data was cleaned and transformed, multiple machine learning algorithms were trained and compared—Logistic Regression, Random Forest, and XGBoost—using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Once the modeling and evaluation were complete, the paper was structured according to the JETIR guidelines, with clearly defined sections including Abstract, Introduction, Methodology, Results, and Conclusion. Tables and evaluation charts were added, and the document was reviewed thoroughly to maintain clarity and alignment with academic standards.

2. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they appear in the text, even if already mentioned in the abstract. Avoid using abbreviations in section headings or the paper title unless absolutely necessary. Below are the abbreviations and acronyms used in this paper:

- **ML** – Machine Learning
- **NLP** – Natural Language Processing
- **TF-IDF** – Term Frequency–Inverse Document Frequency
- **LR** – Logistic Regression
- **RF** – Random Forest
- **XGB** – XGBoost
- **EDA** – Exploratory Data Analysis
- **API** – Application Programming Interface
- **CSV** – Comma-Separated Values
- **ROC** – Receiver Operating Characteristic
- **AUC** – Area Under the Curve
- **TP** – True Positive
- **FP** – False Positive
- **FN** – False Negative
- **TN** – True Negative

III. RESEARCH METHODOLOGY

This section outlines the approach adopted to develop and evaluate a machine learning model for the classification of fake job postings. The methodology involves data collection, preprocessing, feature engineering, model selection, training, evaluation, and validation. The goal is to build a predictive system capable of distinguishing between genuine and fraudulent job advertisements based on textual and categorical data.

3.1 Population and Sample

The population for this study consists of job advertisements posted on online job portals, including both legitimate and fraudulent listings. With the rising use of digital platforms for recruitment, job seekers increasingly rely on these sources, making them vulnerable to scams. The dataset used reflects this real-world scenario, representing a broad cross-section of job postings across various sectors, roles, geographic regions, and industries. The aim is to capture diverse characteristics of both real and fake job postings to enable the machine learning model to generalize effectively.

A sample of 17,880 job postings was obtained from the publicly available Kaggle dataset titled "Fake Job Postings." Out of this total, approximately 3,000 entries were labeled as fake, while the rest were classified as genuine. The sample includes a variety of job types and structures, ranging from entry-level to executive roles. The dataset was cleaned to remove incomplete or ambiguous entries, and stratified sampling was applied to maintain a balanced representation of both classes during training and evaluation. This ensures that the classification models are trained on relevant and diverse examples, enhancing their ability to detect fraudulent patterns.

3.2 Data and Sources of Data

The data used in this study was sourced from publicly available datasets related to job postings, specifically from Kaggle's *Fake Job Postings Dataset*. This dataset contains anonymized job listings and their associated attributes, including job titles, descriptions, company profiles, locations, and employment types. Key features extracted and used for modeling include:

- **Job Title** (e.g., Software Engineer, Data Scientist)
- **Location** (e.g., New York, Remote)
- **Company Profile** (e.g., Company description, size)
- **Job Description** (Textual field outlining job responsibilities)
- **Job Requirements** (e.g., education, skills, experience)
- **Employment Type** (e.g., Full-time, Part-time, Contract)
- **Industry** (e.g., IT, Healthcare, Marketing)
- **Function** (e.g., Engineering, Sales, HR)
- **Benefits** (e.g., Health Insurance, Paid Time Off)
- **Fraudulent** (Target variable: 0 = Real job, 1 = Fake job)

The dataset underwent several preprocessing steps to ensure the quality of the data used for training the model. Textual features, such as job descriptions and company profiles, were tokenized, lemmatized, and vectorized using TF-IDF (Term Frequency-Inverse Document Frequency). Categorical variables, such as employment type and location, were encoded using label and one-hot encoding methods. Continuous features, such as job titles and benefits, were normalized for consistency. Any missing or erroneous data was imputed or removed to avoid compromising model performance, ensuring that the training data remains clean and unbiased.

3.3 Theoretical framework

The objective of this study is to build predictive models that estimate the likelihood of a job posting being fake based on various textual and categorical attributes. The dependent variable is whether the job posting is fraudulent (binary classification: 0 = Real job, 1 = Fake job), while the independent variables include:

- **Title Length** (numerical: number of words in the job title)
- **Company Profile Length** (numerical: number of words in the company description)
- **Description Length** (numerical: number of words in the job description)
- **Requirements Length** (numerical: number of words in the requirements field)
- **Employment Type** (categorical: Full-time, Part-time, Contract, Temporary, Internship)
- **Location** (categorical: e.g., specific city, "Remote")
- **Industry** (categorical: e.g., IT, Healthcare, Marketing)
- **Function** (categorical: e.g., Engineering, Sales, Human Resources)
- **Benefits Mentioned** (binary: 0 = No benefits listed, 1 = Benefits listed)
- **Fraudulent** (target variable: 0 = Real, 1 = Fake)

The relationship between these features and the likelihood of fraud is inherently complex and potentially non-linear, as deceptive postings may mimic legitimate ones in subtle ways. This complexity motivates the use of machine learning algorithms capable of capturing both linear and non-linear patterns. Logistic Regression (LR) provides a benchmark linear model for interpretability, while Random Forest (RF) leverages an ensemble of decision trees to model interactions and reduce overfitting. XGBoost (XGB), a high-performance gradient-boosting method, further enhances predictive accuracy on imbalanced and high-dimensional datasets. Together, these models test the hypothesis that fake job postings exhibit discernible lexical and structural signatures that can be learned and generalized to unseen data.

3.4 Statistical tools and econometric models

For this study, several machine learning algorithms were employed to build predictive models, including Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB). These models were selected for their ability to handle both structured and unstructured data, capturing complex patterns in job posting features. Model performance was evaluated using key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to ensure robust classification of real versus fake job postings.

3.4.1 Descriptive Statistics

The analysis revealed that description lengths ranged from 10 to 800 words (mean = 210, SD = 65), while title lengths varied from 2 to 15 words (mean = 7.9, SD = 2.8). Requirements length spanned from 5 to 200 words (mean = 45, SD = 20). The target variable, "Fraudulent," had 17% fake job postings, highlighting a class imbalance in the dataset.

3.4.2 Machine Learning Models Used

We used and compared several machine learning models to predict whether a job posting is fake or real:

a) Logistic Regression (Baseline Model)

This linear model serves as a baseline, offering interpretability while predicting the probability that a job post is fraudulent based on input features..

b) Random Forest Classifier

An ensemble model that combines multiple decision trees to improve prediction accuracy and handle complex feature interactions while reducing the risk of overfitting

.

c) XGBoost Classifier

A high-performance gradient boosting algorithm that excels in accuracy, speed, and robustness, especially suitable for large, imbalanced, or complex datasets.

d) K-Nearest Neighbour

Predicting brain strokes often involves analyzing medical data using algorithms like K-Nearest Neighbour (KNN) or other machine learning models. These predictions take factors such as age, blood pressure, cholesterol levels, heart rate, and lifestyle habits into account. Models can classify a person into risk categories by comparing their data to existing cases.

3.4.3 Evaluation Metrics

We used the following metrics to evaluate how well each model classified job postings as fake or real:

- **Accuracy**: The percentage of total correct predictions.
- **Precision**: The proportion of correctly predicted fake jobs out of all predicted fake jobs.
- **Recall**: The proportion of correctly predicted fake jobs out of all actual fake jobs.
- **F1-Score**: The harmonic mean of precision and recall, balancing both.
- **ROC-AUC**: Measures how well the model distinguishes between fake and real job postings.

3.4.4 Comparison of the Models

The models were compared based on their evaluation metrics, with special emphasis on F1-Score and ROC-AUC due to the imbalanced nature of the dataset. Among the models tested, Random Forest and XGBoost achieved the highest scores in precision and recall. Feature importance analysis revealed that fields such as job description length, company profile content, and employment type were among the most significant indicators in predicting fraudulent listings.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Descriptive Statics

Variable	Minimum	Maximum	Mean	Std. Deviation
TitleLength	2	15	7.9	2.8
DescriptionLength	10	800	210	65.0
RequirementsLength	5	200	45.0	20.0

Table 4.1 summarizes the statistical characteristics of select features, such as title length, description length, and company profile length. These metrics provide insight into the structure and composition of the job postings and support feature engineering decisions in the modelingphase.

IV. ACKNOWLEDGMENT

The author wishes to express sincere gratitude to the Project Guide and Head of the Department of MCA, Dr. Shashidhar Kini K, for his invaluable guidance, constant encouragement, and kind support throughout this research work. Appreciation is also extended to the Principal, Dr. Shrinivasa Mayya D, for fostering an environment conducive to completing this project within the institution. The author thanks the management of Srinivas Institute of Technology for their direct and indirect support. Gratitude is also due to all the faculty members and non-teaching staff of the MCA department for their constant help and support. Finally, the author is indebted to parents and friends for their unwavering support and belief throughout this endeavor.

REFERENCES

[1] Shivam Bansal. (2019). *Fake Job Postings Dataset*. Kaggle. https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction

[2] R. Patil and A. Shah. (2021). "Detecting Fake Job Postings Using Machine Learning Techniques," *International Journal of Computer Applications (IJCA)*, vol. 183, no. 45, pp. 10–15.

[3] S. Sharma and M. Gupta. (2022). "A Comparative Analysis of Machine Learning Models for Fake Job Detection," *International Journal of Data Science and Analytics*, vol. 9, no. 2, pp. 25–33.

[4] K. Reddy and D. Singh. (2020). "Natural Language Processing Techniques in Job Post Classification," *International Journal of Artificial Intelligence Research*, vol. 14, no. 1, pp. 56–64.

[5] Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

[6] Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830.