International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Fake Job Post Detection Using Machine Learning

Rama Lakshmi[1], N. Chaitanya Sarvani[2], Madhappavar Nikhil[3], Ginjala Meghana[4]

[1,2,3,4]Computer Science and Engineering Matrusri Engineering College Hyderabad, India

Abstract

With online job portals, job searching has become more convenient and accessible, with a tangible increase in fraudulent job posts. A significant part of these listings is well developed, to look authentic and users may struggle to differentiate between authentic opportunities and fraud.

This work has created a machine learning-based system by detecting fake job postings through textual features of the job title, description, and requirements. TF-IDF vectoriza-tion processes the text data to convert it into numbers allowing the models to detect meaningful patterns in the job descriptions. Numerous machine learning models are trained and evaluated to identify the best approach in classification.

The highest performance in the Support Vector Machine (SVM) model was recorded in the course of experimentation in terms of accuracy. The chosen model is implemented in a web application based on Flask, so the user can examine job approvals in real-time and get instant feedback.

Besides prediction, the system also has some practical functions, including scam reporting, scam knowledge base, and one that checks the legitimacy of the company. These attributes make the system useful and offer users with more means of verifying job posts. The findings prove that the system is able to detect suspicious job posts efficiently without being complex and inaccessible to the average users.

Keywords: Fake job detection, machine learning, natural language processing, TF-IDF, text classification, fraud detection, Flask, web application.

INTRODUCTION

Job portals through the Internet have come to be one of the most popular job search platforms. As much as they offer convenient access to opportunities, they have also have led to increased cases of fraudulent postings of jobs [2]. There are numerous forgeries developing around job advertisements in order to be as much similar to gen-uine as possible. Such posts usually come with descriptions of their work, good pay packages and work flexibility. There are other instances where they have been employed in an attempt to obtain sensitive data or to pays people in the guise of registration or training. Due to this reason, job seekers, particularly new graduates and less skilled user are more susceptible to the scam. Older techniques to identify fraudulent job postings are manual verification and rule-based systems. These methods may be effective to certain degree, however, they cannot be used when dealing with huge amounts of data. Moreover, rule-based systems fail to evolve with new and emerging trends in fraud hence restricting their capability in the long term.

A more reliable and easy to scale solution is provided by machine learning. Through training on textual information, it is possible to learn patterns that can be used to dis-tinguish between legitimate and fraudulent job advertisements [1], [8]. These models, in combination with natural language processing techniques can extract meaningful features of job descriptions and classify better.

This work proposes a system that will help categorize the job posters as real or fake through a machine learning algorithm. The system applies TF-IDF to vectorize text to numbers and prediction is obtained by various classification models. In our implementation, the emphasis is on textual analysis of patterns of job descriptions to identify possible fraud. The efficiency of these models is measured in order to provide the most appropriate approach.

The system is operationalized as a web-based app with Flask to make it more practical. The app enables users to enter job information and get predictions instantly. The system also comprises prediction modules, scam reporting, scam library, company checker, an assistant module among others, that assist the users checking on job postings and making informed decisions.

The key aim of creating this work is to develop a system that not only results in good level of prediction accuracy but also a user friendly system that can be used in the real world by means of convenient features. As the trend of relying on online platforms to job seekers continues to increase, systems to assist users to recognize risks are evident. Combining prediction with practical tools can provide a better support than standalone models.

PROBLEM STATEMENT

The rise in the online job portals has contributed to more fraudulent job posts. A lot of these counterfeit listings have been created to closely resemble real jobs and these fake pages posing as jobs cannot be detected by users. Consequently, job applicants face potential risks of losing money and falling into abusers of their data.

Current methods, such as manual confirmation and rule-oriented systems, are unproductive with big data. Such techniques also find it hard to identify emerging and changing patterns of frauds, and this minimizes their usefulness.

To develop a system where the process of identifying job posts as either real or fake by the use of machine learning methods is achieved automatically is the problem used in this work. There should also be assisting features of the system that assist users in gaining a better understanding of and checking job postings.

LITERATURE SURVEY

The need to detect fraudulent job posting has been in the limelight with the growing popularity of online recruitment websites. Loss of money and abuse of personal data are some of the severe consequences of fake job postings that has motivated studies in this field.

In the early years, the conventional machine learn-ing methods that were examined predominantly included the Logistic Regression, Support Vector Machines (SVM) and Decision Trees. These are based on the manually engineered features and would need a lot of textual pre-processing of the text. Although they are able to obtain respectable accuracy, they tend to perform poorly with more complex datasets [2].

Text-based analysis is now more effective with the progress in the field of natural language processing (NLP). The most common methods of transforming text into numbers include Bag-of-Words and TF-

IDF. These strategies can be used to find critical words and patterns in job descriptions and specifications to aid the classification work [6].

The recent studies have examined ensemble learning algorithms like XGBoost and Random Forest. Such models are able to combine various decision trees and are capable of capturing the complex relationship within the data, resulting in better performance than individual models [4]. They come in handy especially when dealing with large and different datasets.

Moreover, there are deep learning methods e.g. Long Short-Term Memory (LSTM) networks that have been used in text classification problems. These modeling types can identify patterns of sequence in written word and have reported positive outcomes in the disclosure of deceptive employment advertisements [8]. Transformer-based mod-els like BERT have shown good performance in comprehending contextual relationships in textual material as well [10].

Although this has occurred, most of the existing systems concentrate mainly on enhancing accuracy in predictions and do not consider usability. Users would need other assistance like verification tools and reporting features in reality.

The proposed system in this work is a combination of machine learning-based prediction and a web-based interface together with some other modules like scam reporting, scam library and company verification. It enhances the accuracy and usability of the system, allowing it to be more real-world applicable.

METHODOLOGY

A. *System Overview*

The suggested system will be in the form of a full pipeline, integrating machine learning algorithms with a web-based interface to identify fake job postings. The system takes the textual data related to a job, applies feature extraction and trains the model to categorize the postings into real and fake. This work is based on usability in contrast to classification only systems since it incorporates this model in a web application. Other modules like scam reporting, scam library, company checker as well as assistant module are incorporated to increase the interaction of the user as well as to give practical help.

B. *Dataset Description*

The dataset of job postings used in this work is the set of job postings obtained in publicly available sources. Every entry has textual entries which include job title, description, and requirements and this would be critical in finding patterns amongst jobs offered.

The data is marked with one binary variable that is whether a job advert is authentic or fake. This ensures that it is appropriate under supervised learning methods whereby the model can acquire distinguishing features.

TABLE I
DATASET FEATURES DESCRIPTION

Sl	Feature	Description
1	title	Job title of the posting
2	description	Detailed job description
3	requirements	Skills and qualifications required
4	company profile	Information about the company
5	fraudulent	Target label (0: real, 1: fake)

C. *Data Preprocessing*

The raw data in the dataset are text data and must be cleaned and then used in training. This includes cleaning up of special characters, lowercasing of text and also dealing with missing values.

Unnecessary symbols and formatting errors are also part of the text preprocessing. Such measures enhance the quality of the information and assist the model to learn better. It is essential to preprocess properly since inconsistencies in the text even in small amounts can influence the performance of the model.

D. *Feature Extraction*

TF-IDF vectors are utilized in order to transform textual data into a form that can be used in machine learning algorithms. This method weight words according to their significance in the dataset.

TF-IDF assists in highlighting more relevant words, whilst minimizing the effect of repetitive words which might not help in classification. This enables the model to concentrate on purposeful patterns that occur in job descriptions.

Fig. 1. Proposed System Architecture

E. *Model Training and Evaluation*

Numerous machine learning models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost are trained and tested. The choice of these models was based on the fact that they are generally applied in text classification processes and they perform well.

Both models are fitted to the processes dataset and analyzed in terms of accuracy. Comparison will be done on various models to determine the best model to use on the problem. SVM performed well in this work than the other models.

Comparison of various models will also be used to interpret the behaviour of various algorithms with respect to text data. This analogy offers greater understanding concerning model selection in real-world implementation. In testing, we found out that various models manifested differently based on the job description structure and length.

F. *Web Application Integration*

Integration of the trained model into a Web app based on Flask is implemented. It enables this to be done in real time allowing the user to apply the job details to get predictions.

The interface is supposed to be user friendly and easy to use by users having varying level of technical expertise. The system displays its results clearly i.e. risk level and scam type.

G. *Additional Modules*

Besides prediction, the system has a number of support modules:

- Dashboard: Gives a summary of features of the system.
- Scam Reporting: Provides possibility to report suspicious job posting.
- Scam Library: Scams that have been reported before.
- Company Checker: Aids in checking company authenticity.

- Assistant Module: Additional support and help to users.

H. *Workflow*

The workflow starts with user input, proceeds to preprocessing and feature extraction. The trained model is then inputted with the processed data in order to be predicted. The result is presented to the user, together with other details.

SYSTEM IMPLEMENTATION

The system will be applied as a web-based application on Flask framework in which the trained machine learning model will run in prediction in real time. Its implementation relates the model at the back-end to a simple front-end interface enabling users to engage with the system without any technical knowledge.

This workflow starts by the user providing job related information such as title, description and requirements via the inter-face. These inputs are preliminarily integrated and pre-processed using the same preprocessing pipeline as during training. This guarantees training and prediction consistency.

The processed input is processed by the saved TF-IDF vectorizer. This feature vector is then sent to the trained model which then makes a prediction as to whether the job posting is genuine or a fake one. The output is then translated into an easily understandable format which contains the level of risk and simple description.

There are other modules in the application, which go further than prediction. The scam reporting tool will enable one to report on suspicious job advert and this will subsequently be verified. The scam library hosts a list of familiar scam patterns, which assists its users to become familiar with the types of frauds.

Company checker module gives an easy option of the company user who is just trying to know whether a company is legitimate or not, given the information he has. The assistant module is user-friendly whereby it suggests and gives precautions during job application.

In implementation, it can be ensured that the machine learning pipeline is separate from the web interface to ensure that the interface remains modular. This facilitates the process of updating the model, as well as making an expansion of the system in future without its impact on other components.

In general, the implementation is based on simplicity, regularity and usability, as well as is successful such that the system is practically being utilized by the job seekers.

RESULTS AND DISCUSSIONS

The test of the proposed system was done with different machine learning models trained on the processed job postings data. TF-IDF vectorization transformed the textual data into numerical features, enabling important patterns of job description to be characterized by the models.

A number of models were experimented upon and these included: Logistic Regression, Support Vector Machine (SVM), Random Forest and XG-Boost. The most accurate model was the Support Vector Machine (SVM) model with a high accuracy of around 97.9. This implies that the model can be effectively used to differentiate actual and fake job advertisements in the provided dataset. The other models worked well as well though the difference in the accuracy was minimal.

A. *Model Performance Summary*

Table II provides an overview of the performance of various machine learning models.

The comparison has shown that SVM is the most accurate and as such, it is chosen as the final model to be deployed.

TABLE II
MODEL PERFORMANCE COMPARISON

Model	Accuracy (%)
Logistic Regression	96.8
Support Vector Machine (SVM)	97.9
Random Forest	97.8
XGBoost	97.7

When using the model during evaluation, it was noted that the model would work well when it came to job postings where the description of the job was detailed enough. With all his efforts, the model was able to categorize structured job features with specific tasks and requirements more easily. Conversely, we have seen occasionally longer or less specific job de-scriptions, which prompted less certain predictions, pointing to the value of input quality in text-based classification.

The other major finding was that some of the illegal postings of jobs resemble exactly what the genuine listings look like. These postings are usually written with professional language and formatting and it is hard to spot them just by the characteristics of the text. Sometimes therefore, the model does not classify the input correctly.

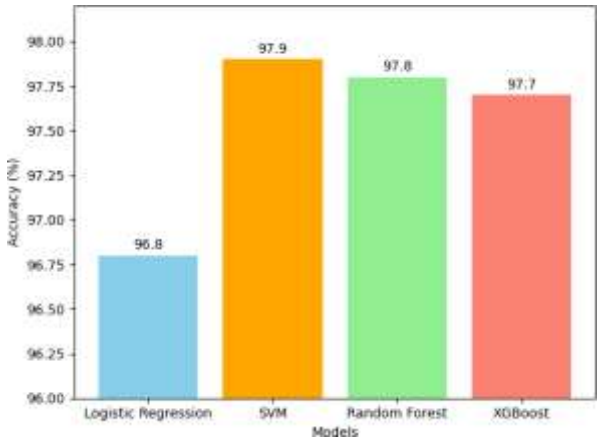

Fig. 2. Accuracy comparison of different machine learning models

It was a Flask-based web application that incorporated the trained model into one that could test the model in real time. Users are able to enter employment information as the title, description and requirements and the system gives a prediction at once. The output is already in a form that is easy to read and understand such as the risk level and a small description, which enhances interpretability.

Several sample job descrip-tions were utilized: valid and suspicious. This prediction was in accordance with the expectations in most cases. The system could accurately detect typical scam ele-ments like unrealistic pay adver-saries, little to no requirements, and advance payments.

The use-fulness of the system is further boosted with inclusion of other modules like scam reporting, scam library and company checker. These aspects enable the users to get beyond prediction and actually confirm job postings in a variety of methods.

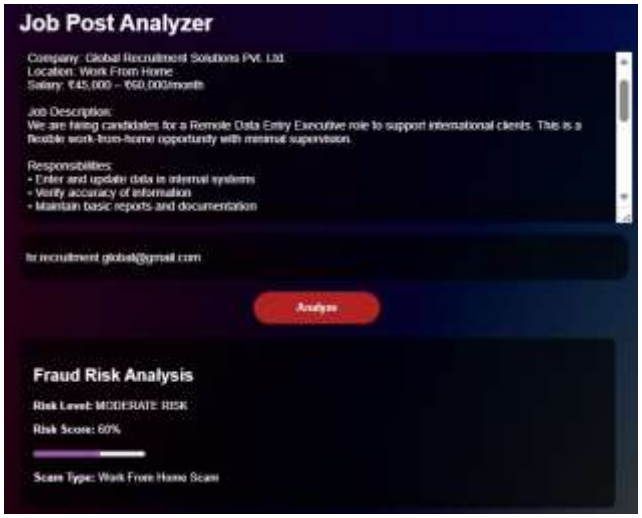

Fig. 3. Web Application Interface for Fake Job Detection

B. *Discussion*

The system-wide results demonstrate that machine learning can be utilized successfully in discovering possibly fraudulent job adverts. TF-IDF vectorization and the use of classification models is an effective method that can be used to analyze text job information.

The system however remains reliant on the quality and extent of diversity of the dataset. Given that the model is mostly based on textual characteristics, writing style differences and lack of job description can influence accuracy of prediction.

The second weakness lies in the fact that there are also very advanced fake postings, which closely similar those that are around. Such instances demand superior methods or extra validation methods to detect the cases.

Regardless of these limitations, the system works fine in the majority of real-life situations and renders a helpful aid to job seekers. The use of bigger datasets, improved models, and enhanced explanation mechanisms could be utilized in the future to further improve performance and usability.

CONCLUSION

In this work, there was the development of a machine learning-based system to detect fake job postings through textual analysis. The system can detect patterns distinguishing real job postings and fake ones by using TF-IDF feature extraction with several classification models.

The system will be more usable and reachable through incorporation of the trained model into a web-based application. The users are able to generate job details whereby the system would give a prediction with some basic descriptions thus making users understand their results better.

This system can be compared with other systems that are based on a single aspect, which is classification; however, it has other elements, highlighted by scam reporting, scam library and verification of the company. These capabilities render the system more applicable in real world scenarios, where users require more features than a prediction.

Despite the high performance of the system in most circumstances, some shortcomings were noted especially when it comes to job postings where virtually no or insufficient information is provided. This means that more improvement is necessary with bigger and more varied datasets.

Further research on improving the performance of the model through high-end methods and broadening

the data can be used in the future. The effectiveness of the system and its usability can also be enhanced by enhancing the mechanism of explanation and adding real-time verification of the verification process.

REFERENCES

1. A. S. Pillai, "Detecting Fake Job Postings Using Bidirectional LSTM," *IRJMETS*, vol. 5, no. 3, 2023.
2. K. Sridevi et al., "Real or Fake Job Posting Detection," *IRJAEM*, 2024.
3. V. Itnal et al., "Fake/Real Job Posting Detection Using Machine Learn-ing," *IJRASET*, 2025.
4. T. N. Goud and R. Reddy, "A Machine Learning Approach for Detecting Fraudulent Job Postings," *Journal of Sensors IoT Health Sciences*, 2025.
5. S. Ramya et al., "Fake Job Prediction Using Machine Learning Algo-rithms," *IJERST*, 2025.
6. M. Kumari et al., "Fake Job Posting Prediction Using Machine Learning Approach," *IJERT*, 2023.
7. P. G. B. et al., "Fake Job Detection Using Hybrid Network-Based Approach," *International Journal of Environmental Sciences*, 2025.
8. E. Baraneetharan, "Detection of Fake Job Advertisements Using Ma-chine Learning Algorithms," *Journal of Artificial Intelligence and Cap-sule Networks*, 2022.
9. Pradeep et al., "Fake Job Posting Detection Using NLP and XGBoost," *IJRASET*, 2025.
10. Jayesh Fating et al., "Fake Job Listing Detection Using Machine Learning," *IJRASET*, 2023.
11. P. Hongal et al., "Fake Job Detection Using Machine Learning," *IJPREMS*, 2024.
12. Zhang et al., "Deep Learning Approaches for Job Fraud Detection," 2021.
13. Chen et al., "Machine Learning Classification of Job Postings," 2018.
14. T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," 2013.
15. J. Ramos, "Using TF-IDF for Information Retrieval," 2003.
16. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD*, 2016.
17. J. Lee and M. Cho, "Online Job Scams: Detecting Fraudulent Job Advertisements Using Machine Learning," *New Media & Society*, 2025.
18. V. Itnal et al., "Fake/Real Job Posting Detection Using Machine Learn-ing," *IJRASET*, vol. 13, no. 10, 2025.
19. M. Krishna et al., "Unmasking Employment Scam: Automated Detection of Fraudulent Job Postings," *Global Journal of Engineering Innovations*, 2025.
20. "Fake Job Detection: A Comparative Study of Machine Learning and Hybrid Approaches," *International Journal of Environmental Sciences*, 2025.