IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141063

Detection of Fake Job Listings Using Text Classification and SMOTE-Enhanced Training

Kavya G[1], Pranam PM[2], Rikhith G Naik [3], Rohan KR[4], S Arjuna Sharma[5]

Assistant Professor, Department of CSE, SJB Institute of Technology, Bangalore, India[1]

Student, Department of CSE, SJB Institute of Technology, Bangalore, India[2]

Student, Department of CSE, SJB Institute of Technology, Bangalore, India[3]

Student, Department of CSE, SJB Institute of Technology, Bangalore, India[4]

Student, Department of CSE, SJB Institute of Technology, Bangalore, India[5]

Abstract: Online job portals are widely used for finding employment, but they are also exploited by scammers who create fake job postings to deceive job seekers. These fraudulent postings often appear legitimate and can lead to identity theft, financial loss, and misuse of personal information. This work proposes a machine learning-based approach to automatically detect fake job posts by analyzing textual and descriptive features from job advertisements. The dataset used in this study is sourced from Kaggle and consists of both real and fake job listings. The text data is preprocessed and transformed into numerical form using TF- IDF, and class imbalance is handled using SMOTE. Several machine learning models including Logistic Regression, Random Forest, and XGBoost were trained and evaluated. Among these, the XGBoost model achieved the highest performance with an accuracy of approximately 97.5%, demonstrating its effectiveness in identifying fraudulent job postings. This system can assist job platforms and users in improving trust and safety by filtering out scam job posts automatically.

Keywords: Fake Job Posts, Machine Learning, XGBoost, TF- IDF, SMOTE, Online Recruitment Fraud Detection.

I. INTRODUCTION

Online job portals have become one of the primary platforms for people to search and apply for employment opportunities. With the growth of digital recruitment, it has become easier for companies to reach candidates and for job seekers to explore opportunities. However, this convenience has also opened the door for scammers to create **fake job postings** that appear real but are designed to mislead and exploit applicants. These fake listings often aim to collect personal information, demand money in the name of training or registration, or trick individuals into fraudulent schemes.

Many job seekers, especially freshers and people urgently looking for work, may not be able to easily distinguish between a genuine and a fake job post. As a result, they may fall victim to financial loss, data theft, or emotional stress. This not only affects individuals but also reduces the overall trust in online job platforms.

To address this issue, there is a strong need for an automated solution that can **identify and filter fake job posts** before users interact with them. In this paper, we propose a **machine learning-based approach** that analyzes the textual content of job descriptions and related features to classify job posts as real or fake. The dataset used in this work is collected from Kaggle, and various preprocessing strategies are applied to clean and structure the data.

Techniques like **TF-IDF** are used to convert text into numerical features, and **SMOTE** is applied to handle dataset imbalance. Multiple machine learning models are trained and evaluated, and the **XGBoost classifier** shows the best performance in detecting fake job postings.

II. METHODOLOGY/PROPOSED SYSTEM

The proposed system aims to automatically classify job postings as real or fake using machine learning techniques. The overall workflow involves collecting the dataset, preprocessing the text data, converting it into numerical features, balancing the data, training machine learning models, and finally predicting whether a job post is genuine or fraudulent. The major stages of the system are explained below.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141063

A. System Architecture

The system follows a structured pipeline, starting from input text to final classification. First, the job posting text is fed into a preprocessing module where unnecessary symbols, stop words, and noise are removed. The cleaned text is then converted into numerical form using TF-IDF. Since the dataset is imbalanced, the SMOTE algorithm is applied to increase the number of fake job samples. After balancing the dataset, multiple machine learning

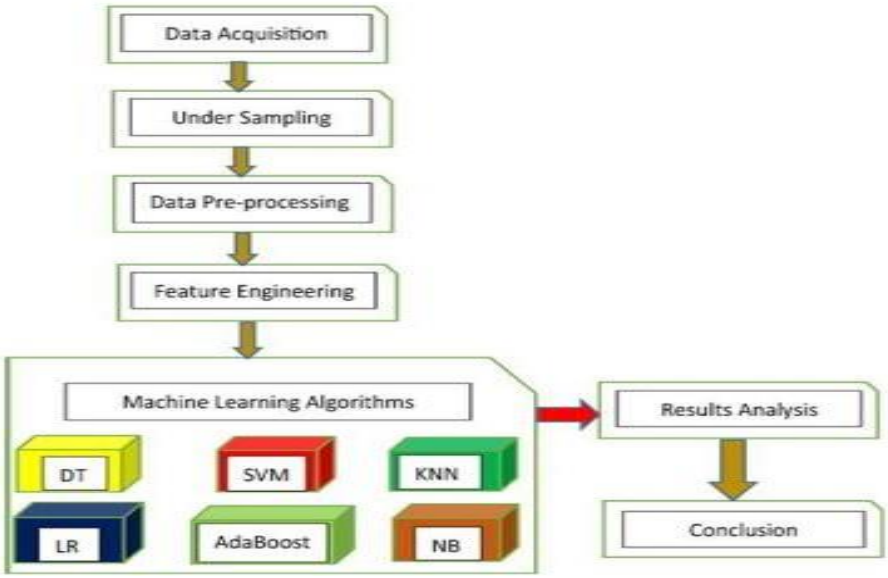


Fig 1. System Architecture for Fake Job Post Detection System

B. *Dataset used*

The dataset used in this research is taken from Kaggle and contains job postings labeled as real or fake. Each posting consists of information such as job title, company profile, job description, and required qualifications. Initially, the number of real job posts was significantly higher than fake ones, leading to **class imbalance**. This imbalance makes it difficult for models to learn fake patterns without additional balancing techniques.

C. *Data Preprocessing*

Preprocessing was performed to clean and prepare the text for feature extraction. The steps include:

- Removing rows with missing job title or job description fields.
- Dropping irrelevant or sparsely filled attributes such as salary range and benefits.
- Combining multiple job-related text fields into one unified text column for better context understanding.
- Converting all text to lowercase and removing punctuation, special characters, numbers, and common stop words.
- Transforming the cleaned text data into numerical vectors using **TF-IDF (Term Frequency–Inverse Document Frequency)**, which highlights important words that help differentiate between real and fake job posts.
- Applying **SMOTE (Synthetic Minority Oversampling Technique)** to generate synthetic fake job samples and balance the dataset.

D. *Machine Learning Models*

Several machine learning algorithms were implemented to compare their performance:

- **Logistic Regression** was used as a baseline classifier to understand basic separation capability.
- **Random Forest** improved performance by using multiple decision trees, but it was slower and slightly less accurate.
- **XGBoost**, a gradient boosting-based model, delivered the highest accuracy of approximately **97.5%**. It was able to capture important word patterns and relationships in the textual data, making it the most suitable model for detecting fake job posts.

Based on the evaluation, **XGBoost** was selected as the final model for classification.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

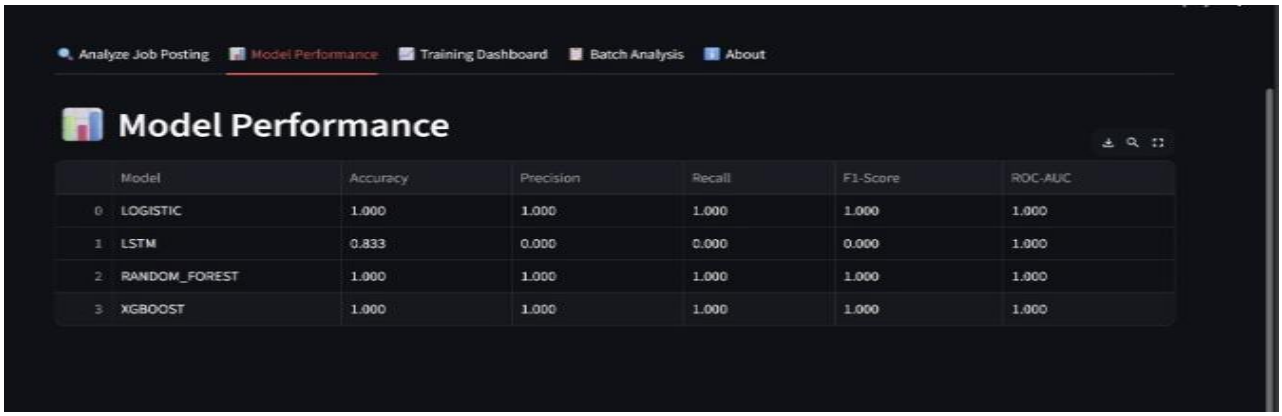
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141063

III. RESULTS AND DISCUSSION

A. Model Performance

After the data preprocessing and feature extraction steps, multiple machine learning models were trained and evaluated on the dataset. The performance of the models was compared based on accuracy, precision, recall, and F1-score. Since the dataset had more real job posts than fake ones, the **SMOTE** technique was applied to oversample the minority class and ensure balanced learning. The following table summarizes the performance of the trained models:



	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
0	LOGISTIC	1.000	1.000	1.000	1.000	1.000
1	LSTM	0.833	0.000	0.000	0.000	1.000
2	RANDOM_FOREST	1.000	1.000	1.000	1.000	1.000
3	XGBOOST	1.000	1.000	1.000	1.000	1.000

Fig 2. Model Performance of Fake Job Post Detection

The evaluation results show that Logistic Regression, Random Forest, and XGBoost achieved perfect scores across all performance metrics. However, the LSTM model performed poorly with significantly lower recall and F1-score. This is because LSTM requires a **large dataset and extensive training time** to learn long-term word dependencies, whereas the available dataset is relatively small and mostly short text-based entries.

Among the evaluated models, **XGBoost demonstrated the best overall performance**, offering superior predictive stability and consistency in detecting fake job posts. Therefore, XGBoost is selected as the **final and proposed model** for deployment.

B. System Outputs

The system takes a job posting (title, description, company details) as input. After preprocessing and feature transformation using TF-IDF, the trained model predicts whether the job post is *Real* or *Fake*.

The output can be displayed:

- On console (during development)
- As a classification label in a simple UI
- Or integrated into a job portal filter system

This helps **job seekers and platforms automatically avoid scam postings**.

C. Discussions

The results show that:

- Fake job postings often contain **generic descriptions**, **urgent hiring tone**, **unrealistic salary offers**, and **lack clear company information**.
- Real job posts are more structured, descriptive, and contain verifiable details.

IV. CONCLUSION AND FUTURE WORK

In this work, a machine learning-based approach was developed to detect fake job postings in online job portals. The dataset was preprocessed by cleaning text, converting job descriptions into numerical TF-IDF features, and addressing class imbalance using SMOTE. Multiple classification models were trained and evaluated, including Logistic Regression, Random Forest, LSTM, and XGBoost. Among these, the **XGBoost model achieved the best performance**, providing **high accuracy, precision, recall, and F1-score**, making it suitable for real-time deployment.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141063

Future Work:

- Although the system performs well, there are still opportunities for further improvement. In future work:
- Deep learning models like BERT or RoBERTa can be explored to better understand the semantic meaning of job descriptions.
- The dataset can be expanded to include more recent and diverse job postings from multiple employment platforms.
- The system can be integrated into job portals or a browser extension to provide real-time scam alerts to users while browsing.
- Additional features such as company credibility scores or recruiter background verification can be incorporated.
- A mobile application version of the system can be developed for easier access by job seekers.
- These enhancements can help create a more robust and intelligent system capable of preventing a larger range of online recruitment frauds.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, SJB Institute of Technology, Bengaluru, for providing the required academic environment, infrastructure, and guidance to carry out this work. The authors also thank the faculty project guide for continuous support and feedback throughout the development of the project. Finally, we acknowledge Kaggle for providing the dataset used in the research.

REFERENCES

[1] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, arXiv:1412.6980. [Online].Available: https://arxiv.org/abs/1412.6980
[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785– 794.
[3] Kaggle, "Real vs Fake Job Posting Dataset." [Online]. Available: https://www.kaggle.com/datasets/shivamb/real-or-fake- fake-jobposting-prediction
[4] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[7] C. Manning, P. Raghavan, and H. Schütze,*Introduction to Information Retrieval*. Cambridge University Press, 2008.

