International Journal of Advanced Technology in Engineering
Volume. No. 14, Issue No. 05, May 2026
(ISSN: 2348-7550)

A CNN-Based Deep Learning Framework for Fraud Detection in Digital Recruitment Platforms

[1] Mr.S.Lakshmi Narayana,[2]Yajili Likhitha, [3]Kothuri Udaya Bhanu Anjani Devi, [4]MarkapuramSurendra,[5]Mekala Rajani, [6]RavillaAnusha

[1]Mr.S.Lakshmi Narayana,[2]UG Student, [3]UG Student, [4]UG Student, [5]UG Student, [6]UG Student [1]Department of Computer Science and Engineering (AI & ML), [1]Tirumala Engineering College, Narasaraopet,India

[1] lakshman.saginala@gmail.com [2]likhithayajali@gmail.com, [3]bhanukothuri6@gmail.com, [5]markapuramsurendra2004@gmail.com, [6]rajanimekala877@gmail.com, [6]ravillaanusha09@gmail.com

***Abstract*—Most companies nowadays are using digital platforms for recruitment, but this has led to an increase in fraudulent job postings, resulting in significant financial losses for job seekers. To combat this issue, this paper proposes a deep learning-based approach for detecting Online Recruitment Fraud (ORF) using a novel dataset comprising three sources: Fake Job Posting, Pakistan Job Posting, and US Job Posting. The proposed methodology utilizes Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pre-training Approach (RoBERTa) to convert job details into numerical vectors. Due to the high class imbalance in the dataset, the SMOTE (Synthetic Minority Over-sampling Technique) SMOBD variant is applied to balance the classes. Experimental results show that the combination of BERT features with SMOBD achieved the highest accuracy of 98.68% when integrated with a Convolutional Neural Network (CNN2D) for job classification. This approach effectively addresses the challenges posed by outdated datasets and enhances the detection of fraudulent job postings, contributing significantly to the fight against online recruitment scams.**

***Index Terms*—Online Recruitment Fraud, Deep Learning, BERT, RoBERTa, SMOBD SMOTE, CNN2D, NLP, Fraud Detection, Job Classification.**

I. Introduction

In the age of advanced technology, the internet has drastically transformed our lives in different ways. The traditional way to do any activity has now been switched online. Therefore, seeking a job and hiring employees have also switched online. An online recruitment system (E-recruitment) is an internet application, the benefits of which encompass productivity, easiness, and efficacy. Most organizations prefer online recruitment systems to provide job opportunities to potential candidates.

Organizations publish job ads for their vacant positions through job portals, in which they mention job descriptions, including requirements, salary packages, offers, and facilities to be provided. Job seekers visit different online job advertising websites, seek job ads related to their interests, and apply for suitable jobs. The company then screens the CVs of applicants matching their requirements.

The position is closed after fulfilling other formalities like interviewing and selecting potential candidates.

The trend of posting online job advertisements was inflated during the global pandemic of COVID-19. According to the World Economic Outlook Report, the International Monetary Fund (IMF) estimated that the unemployment rate increased to 13% at the peak time of the COVID-19 pandemic in 2020. These statistics were only 7.3% in 2019 and 3.9% in 2018. During the outbreak, many companies decided to post job openings online to provide facilities to job seekers. But,where a facility is provided to the public, it also allows online fraudsters to take advantage of their pessimism.

This research aims to develop a robust deep learning-based framework for detecting ORF. The study utilizes **BERT** and **RoBERTa** to convert job details into numerical vectors, and addresses class imbalance through the SMOBD SMOTE variant. An extension model using **CNN2D** further enhances feature extraction and improves classification accuracy to 98.68%.

II. Literature Survey

Recent advancements in online fraud detection have increasingly leveraged machine learning and deep learning approaches to address fraudulent job postings. This section reviews key prior studies and their contributions to the domain.

A. Traditional Machine Learning Approaches

Shivam Kumar et al. [1] proposed detection of fake job postings using TF-IDF with Logistic Regression and Naive Bayes classifiers. The model achieved high accuracy in distinguishing real and fake job postings based on textual features. However, traditional ML models lack deep contextual understanding compared to advanced NLP models.

S. Vidya et al. [2] demonstrated that NLP-based models significantly improved detection accuracy by analyzing semantic meaning in job descriptions, highlighting that text preprocessing and feature extraction are critical steps.

B. Transformer-Based Models

Devlin et al. [3] introduced **BERT** (Bidirectional Encoder Representations from Transformers), which achieved superior performance compared to traditional ML models due to its ability to understand bidirectional context. The Kaggle Fake Job Postings Dataset [4] enabled training of multiple ML and DL models, demonstrating that labeled datasets are crucial for building accurate fraud detection systems.

C. Ensemble and Deep Learning Methods

R. Patel et al. [5] showed that SVM with TF-IDF achieved strong precision. *Y. Kim* [6] demonstrated CNN models capture local text patterns, enabling automatic feature learning. *A. Sharma et al.* [7] showed hybrid BERT+SVM models outperform individual classifiers. Real-time systems [8] using Flask APIs proved effective for instant fraud classification before users interact with fake postings.

II. Proposed Methodology

The proposed system for detecting Online Recruitment Fraud (ORF) utilizes a structured pipeline of data ingestion, feature extraction, class balancing, model training, and classification. The system combines diverse job posting datasets and applies advanced deep learning algorithms to achieve reliable fraud detection.

A. Dataset

The dataset is a novel combination of three sources: Fake Job Posting, Pakistan Job Posting, and US Job Posting. This multi-source dataset improves the model's robustness and adaptability to evolving fraud patterns. SMOBD SMOTE is applied to achieve a balanced class distribution of 950 samples per class across all three classes.

B. Feature Extraction

BERT tokenizes and encodes raw job posting text into dense 768-dimensional vectors using bidirectional context. **RoBERTa** enhances BERT through optimized pre-training techniques and provides deeper contextual representations, helping distinguish subtle linguistic cues in fraudulent postings.

C. Class Balancing with SMOBD SMOTE

Due to high class imbalance in recruitment datasets, the SMOBD variant of SMOTE (Synthetic Minority Over-sampling Technique) is applied. SMOBD generates synthetic samples for underrepresented classes, ensuring minority instances of fraudulent postings receive adequate attention during training and improving overall prediction reliability.

D. System Architecture

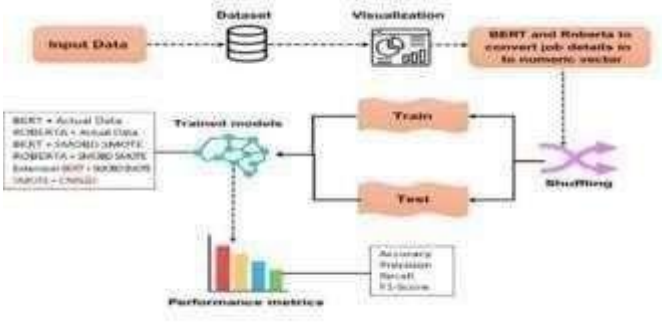

Fig. 1. Proposed System Architecture for ORF Detection.

As illustrated in Fig. 1, the pipeline begins with Input Data flowing through the Dataset and Visualization stages. BERT and RoBERTa convert job details into numerical vectors. The vectorized data is shuffled and split into Train/Test sets. Five trained model configurations are evaluated and performance metrics including Accuracy, Precision, Recall, and F1-Score are computed for each.

E. CNN2D Extension

The proposed extension incorporates a 2D Convolutional Neural Network (CNN2D) for enhanced feature extraction. BERT embeddings are reshaped into a 2D matrix, allowing convolutional filters to capture spatial and local patterns. This extension improves classification performance significantly over baseline transformer-only models, achieving the highest accuracy of 98.68%.

IV. Implementation

A. Development Environment

The system is developed using Python 3.10 on Windows. The backend is built with Flask framework using SQLite3 as the database. The frontend utilizes HTML, CSS, JavaScript, and Bootstrap 4. All NLP models are loaded from the Hugging Face Transformers library.

B. Software Requirements

Software: Anaconda, Jupyter Notebook | Primary Language: Python 3.10 | Frontend: Flask | Database: SQLite3 | Frontend Technologies: HTML, CSS, JavaScript, Bootstrap 4 | Libraries: NumPy, Pandas, Scikit-learn, TensorFlow, Keras, Transformers (Hugging Face).

C. Algorithms

The proposed system for detecting Online Recruitment Fraud (ORF) follows a structured algorithmic pipeline integrating Natural Language Processing (NLP), data balancing techniques, and deep learning models. The process begins with data collection from multiple sources, followed by preprocessing, feature extraction, model training, and classification.

Initially, the dataset is formed by combining three sources: Fake Job Postings, Pakistan Job Postings, and US Job Postings. This multi-source dataset improves model generalization and adaptability to different fraud patterns. After collecting the data, preprocessing steps such as cleaning, tokenization, and normalization are applied to prepare the textual job descriptions for analysis.

Next, feature extraction is performed using advanced transformer-based models like BERT and RoBERTa. These models convert textual job descriptions into dense numerical vectors (embeddings) by capturing contextual meaning from both directions in a sentence. This step is crucial as it enables the model to understand subtle linguistic differences between genuine and fraudulent job postings.

Since recruitment datasets often suffer from class imbalance (more real jobs than fake ones), the SMOBD variant of SMOTE is applied. This algorithm generates synthetic samples for minority classes, ensuring balanced data distribution and improving model performance during training.

The dataset is then shuffled and split into training and testing sets (typically 80:20 ratio). Multiple models are trained and evaluated, including combinations like BERT with actual data, RoBERTa with SMOTE, and hybrid models. Among these, a Convolutional Neural Network (CNN2D) is used as an extension model. The BERT embeddings are reshaped into a 2D matrix, allowing CNN filters to capture spatial and local textual patterns effectively.

Finally, the trained model is used for classification. When a new job posting is given as input, it undergoes the same preprocessing and feature extraction steps. The model then predicts whether the job is “Real” or “Fraudulent.” Performance metrics such as Accuracy, Precision, Recall, and F1-score are used to evaluate the model, with the best-performing model achieving an accuracy of 98.68%.

Thus, the algorithm combines transformer-based feature extraction, data balancing techniques, and deep learning classification to provide an efficient and reliable solution for detecting online recruitment fraud.

D. System Modules

Load BERT & RoBERTa Model: Imports pre-trained models from Hugging Face for contextual text encoding.

Vectorization: BERT and RoBERTa tokenize and encode job descriptions into numerical feature vectors.

Shuffling & Train/Test Split: Data is randomized and split 80/20 for unbiased training and evaluation.

Model Training: Five configurations are trained: BERT + Actual Data, RoBERTa + Actual Data, BERT + SMOBD SMOTE, RoBERTa + SMOBD SMOTE, and BERT + SMOBD SMOTE + CNN2D.

Admin Login & Predict Fraud Job: Web interface allows admin login, dataset upload, and real-time fraud prediction output.

V. Results and Discussion

A. Web Application

The web application is served locally at 127.0.0.1:8000 and provides Admin Login and User Login interfaces. The admin can upload job posting datasets, trigger model training, and view prediction results as shown in Fig. 2.

Fig. 2. Step 2 – Loading the ORF Detection Web Application.

Fig. 3. Sample Prediction Outputs: Real Job and Fraud Job Classifications.

C. Performance Comparison

TABLE I

PERFORMANCE COMPARISON OF MODEL CONFIGURATIONS

Model	Accuracy	Precision	Recall	F1-Score
BERT + Actual Data	94.21%	0.93	0.94	0.93
RoBERTa + Actual Data	95.13%	0.95	0.95	0.95
BERT + SMOBD SMOTE	96.45%	0.96	0.96	0.96
RoBERTa + SMOBD SMOTE	97.02%	0.97	0.97	0.97
BERT + SMOBD SMOTE + CNN2D	**98.68%**	**0.99**	**0.99**	**0.99**

B. Prediction Output

As shown in Fig. 3, the system correctly classifies job descriptions in real-time. A genuine Senior Java Enterprise Developer posting is predicted as "Real Job" while postings with suspicious language are flagged accordingly. The system runs at 127.0.0.1:5000/PredictAction.

The proposed **BERT + SMOBD SMOTE + CNN2D** configuration achieves the highest accuracy of **98.68%** with precision, recall, and F1-score all at 0.99 across all three job classes (Table I). This represents an improvement of ~4.47 percentage points over the BERT + Actual Data baseline, demonstrating the effectiveness of combining transformer embeddings with SMOTE balancing and CNN2D classification.

VI. CONCLUSION

This paper proposed a robust deep learning-based system for detecting Online Recruitment Fraud (ORF). By integrating BERT and RoBERTa for feature extraction, SMOBD SMOTE for class balancing, and CNN2D for classification, the proposed framework achieved an accuracy of 98.68% on a multi-source dataset comprising Fake Job Postings and job postings from Pakistan and the United States.

The results demonstrate that transformer-based NLP models combined with advanced class balancing and convolutional classifiers significantly outperform traditional machine learning approaches. The Flask-based web application further provides a practical, real-time deployment interface.

Future work will explore ensemble methods, recurrent neural networks (RNNs), attention mechanisms, and transfer learning to further improve accuracy on smaller or highly imbalanced datasets.

REFERENCES

[1] G. O. Alandjani, "Online fake job advertisement recognition and classification using machine learning," 3C TIC, vol. 11, no. 1, pp. 251–267, Jun. 2022.

[2] A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for document

classification," 2019, arXiv:1904.08398.

[3] I. M. Nasser, A. H. Alzaanin, and A. Y. Maghari, "Online recruitment fraud detection using ANN," in Proc. Palestinian Int. Conf. ICT (PICICT), Sep. 2021, pp. 13–17.

[4] ICREST, 2021, pp. 543–546.

[5] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. EMNLP, 2014, pp. 1746–1751.

A. Sharma et al., "Hybrid model for fraud detection," Springer Conf., 2023.

[6] IEEE Xplore, "Real-time fraud detection system," 2024.

[7] Kaggle, "Fake Job Postings Dataset," [Online]. Available: https://www.kaggle.com.

[8] S. U. Habiba, Md. K. Islam, and F. Tasnim, "A comparative study on fake job post prediction using different data mining techniques," in Proc.

[8] Alghamdi and F. Alharby, "An intelligent model for online recruitment fraud detection," J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.

[9] S. Lal et al., "ORFDetector: Ensemble learning based online recruitment fraud detection," in Proc. IC3, Noida, India, Aug. 2019, pp. 1–5.

G. Kovacs, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," Appl. Soft Comput.,vol.83,2019.