



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

Online Recruitment Fraud (ORF) Detection using Deep Learning

Dinesh Kumar. V ¹, Yuvaraj. S ², Vignesh Kumar. R³, Vijayendran. S⁴, T.A. Krithika ⁵

B. Tech Students, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India^{1,2,3}

Assistant Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai,
Tamil Nadu, India⁴

ABSTRACT: The rise of digital platforms for job recruitment has brought with it a growing threat of fraudulent job postings, undermining the trust and safety of online hiring systems. This paper proposes an advanced fraud detection system based on the ALBERT (A Lite BERT) model to identify fraudulent job postings. The system will utilize a dataset created by merging job postings from multiple sources to better capture both legitimate and fraudulent job listings. A comprehensive pre-processing pipeline, including data cleaning and feature engineering, will be applied to prepare the data for model training. The system will incorporate various techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), to address class imbalance. The ALBERT model will then be fine-tuned to classify job postings, and the system will be evaluated using standard performance metrics like accuracy, precision, recall, and F1-score. The goal is to provide an effective and scalable solution for detecting fraudulent job postings in real-time, enhancing the overall security and trustworthiness of online recruitment platforms.

KEYWORDS: Online Recruitment Fraud, Deep Learning, ALBERT, SMOTE, Job Posting Classification.

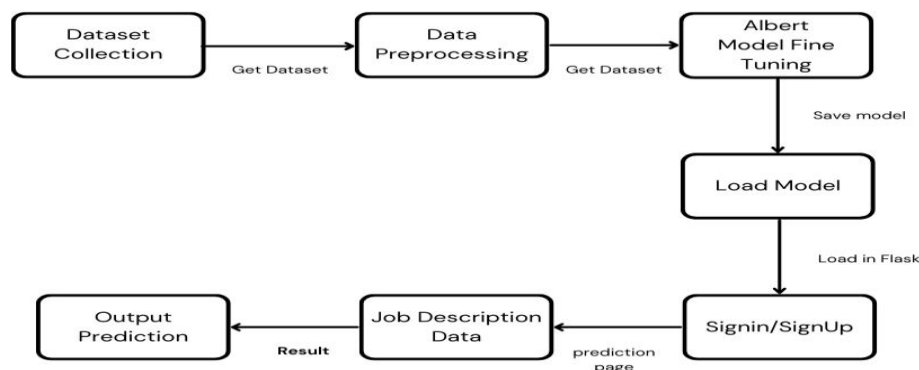


Figure 1: **GENERAL ARCHITECTURE**

I. INTRODUCTION

The project introduces an AI-driven system to address ORF, utilizing the ALBERT model. ALBERT is described as a lightweight version of BERT, chosen to improve fraud detection. Traditional models like BERT and RoBERTa are noted to have limitations, particularly with imbalanced datasets. The approach involves creating a new dataset from three different sources. This new dataset aims to better reflect current trends in online recruitment. The system applies transformer-based models along with advanced balancing strategies. These advanced balancing strategies are used to

improve the classification of job postings. The primary goal is to enhance fraud detection in online recruitment. The research emphasizes overcoming the shortcomings of existing models

II. LITERATURE REVIEW

Online Recruitment Fraud (ORF) has gained attention due to the increasing reliance on digital hiring platforms. Vidros et al. (2017) highlighted the behavioral patterns of fraudulent job postings. Dutta and Bandyopadhyay (2020) employed traditional machine learning models like Decision Trees and SVMs to classify postings. Lal et al. (2019) explored ensemble techniques to enhance prediction accuracy. More recent works introduced deep learning approaches using LSTM and ANN models for better performance (Nasser et al., 2021). However, these models struggled with imbalanced datasets and limited linguistic context. Transformer-based models like BERT showed promise but were computationally intensive. Our study focuses on ALBERT, a lighter alternative, paired with advanced resampling techniques to improve fraud detection accuracy.)

III. METHODOLOGY

To detect online recruitment fraud effectively, a custom dataset was constructed by merging three publicly available job posting sources. The raw data underwent thorough preprocessing steps including removal of null values, normalization of text, and renaming inconsistent column headers. Unstructured text was tokenized using ALBERT's tokenizer for deep contextual understanding. A train-validation split of 80:20 was used to ensure fair evaluation. Additional data analysis was conducted to assess class distribution and detect imbalance. This step was essential to inform the choice of resampling techniques.

The dataset exhibited a significant imbalance between legitimate and fraudulent job posts. To mitigate this, various SMOTE (Synthetic Minority Over-sampling Technique) variants were applied, including SMOTE-NC and SMOBD. These methods synthetically generated new minority class samples to improve model generalization. Feature selection was also performed to retain only the most relevant job description attributes. ALBERT, a transformer-based language model, was then fine-tuned for binary classification using the HuggingFace Transformers library. Hyperparameters were optimized through trial and error for best results..

Model training was conducted using the PyTorch backend with early stopping to prevent overfitting. Evaluation metrics included accuracy, precision, recall, and F1-score to ensure balanced performance. To validate real-world applicability, the trained model was integrated into a Flask-based web application with a MySQL backend. This interface allowed users to input job descriptions and receive fraud predictions instantly. Logging and monitoring features were also added to track predictions. The entire methodology ensured

IV. IMPLEMENTATION

Step 1: Environment Setup

Begin by installing essential libraries including transformers, torch, sklearn, imbalanced-learn, pandas, and Flask. Configure the development environment to support GPU acceleration for faster training.

Step 2: Data Preprocessing

Clean the dataset by handling missing values, normalizing text, removing duplicates, and unifying column names. Tokenize the job descriptions using ALBERT's tokenizer to prepare input for the model.

Step 3: Handling Imbalanced Data

Use SMOTE and its variants (e.g., SMOTE-NC, SMOBD) to generate synthetic examples for the minority class. This ensures the model does not overfit to the dominant class..

Step 4: Model Training

Fine-tune the ALBERT model using a supervised learning approach. Split the dataset into training and validation sets (80:20), and monitor training with metrics like loss, accuracy, and F1-score.

Step 5: Evaluation and Optimization

Evaluate the model using precision, recall, F1-score, and balanced accuracy. Perform hyperparameter tuning to optimize learning rate, batch size, and epochs for best performance.

Step 6: Web Application Integration

Develop a user-friendly Flask interface where users can paste job descriptions. The backend sends the input to the trained ALBERT model and returns real-time fraud predictions.

Step 7: Database and Logging

Integrate a MySQL database to store user queries and results. Implement logging to monitor predictions and collect user feedback for future model improvement.

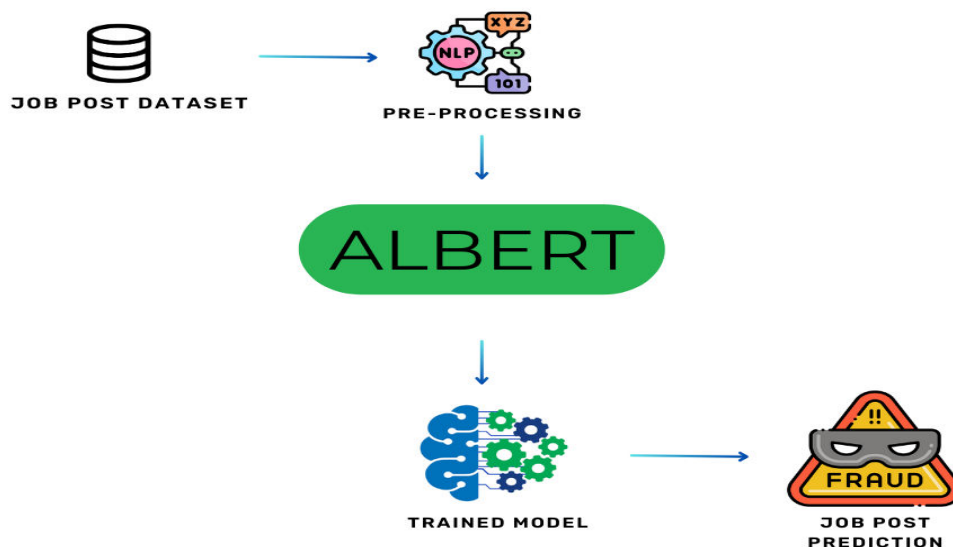


Figure 2: Module Design

The system architecture begins with the user submitting a job description through a web interface. This input is sent to the backend, where preprocessing steps like text cleaning and tokenization using ALBERT tokenizer are applied. The processed text is passed to the fine-tuned ALBERT model, which classifies the job posting as either legitimate or fraudulent. If the model encounters imbalance in classes, SMOTE techniques are used to rebalance the data before training. The classification result is then sent back to the frontend for display. Simultaneously, results and user activity are logged into a MySQL database for tracking. The Flask framework manages communication between components. This architecture ensures secure, scalable, and real-time fraud detection..

V. RESULTS AND CONCLUSION

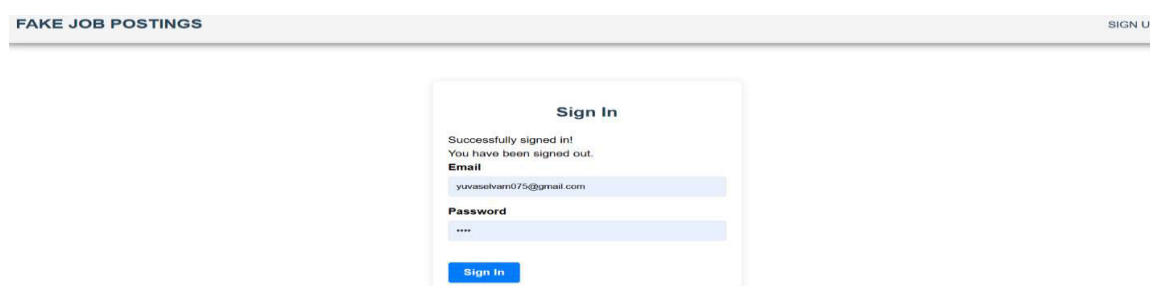
The proposed system effectively classified job postings as fraudulent or legitimate using the ALBERT model in combination with advanced SMOTE techniques. Among all model configurations, ALBERT combined with SMOBD achieved the highest performance, recording approximately 90% balanced accuracy and strong recall, which is crucial for minimizing false negatives. The model outperformed traditional machine learning approaches and even other transformer models like BERT in terms of efficiency and precision on imbalanced data.

The integration of the model into a real-time web application demonstrated its practical viability. Users could input job descriptions and receive accurate predictions instantly, with a smooth and secure interface backed by MySQL. The

results validated that transformer-based models, when combined with proper data balancing and preprocessing, can significantly improve fraud detection in online recruitment.

In conclusion, this research showcases a powerful AI-based solution for detecting ORF. It contributes not only a high-performing model but also a full-stack implementation that can be deployed on real-world platforms. Future work may focus on model interpretability, multilingual support, and expanding the dataset to improve adaptability and trust in diverse environments.

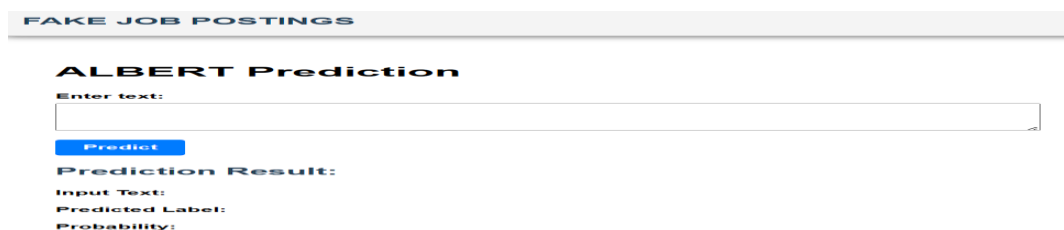
5.1. LOGIN PAGE



The screenshot shows a web application header with "FAKE JOB POSTINGS" on the left and "SIGN UP" on the right. The main content area is titled "Sign In". It contains a message: "Successfully signed in! You have been signed out." Below this, there are input fields for "Email" (containing "yuvaseelvam075@gmail.com") and "Password" (masked with dots). A blue "Sign In" button is at the bottom.

Figure 3: Output

5.1. ALBERT PREDICTION PAGE



The screenshot shows a web application header with "FAKE JOB POSTINGS" on the left. The main content area is titled "ALBERT Prediction". It has a text input field labeled "Enter text:". Below the input field is a blue "Predict" button. Underneath the button, the text "Prediction Result:" is displayed, followed by labels for "Input Text:", "Predicted Label:", and "Probability:", each followed by a blank space for output.

Figure 4: Output

VI. FUTURE WORK

Future improvements include expanding the dataset with real-time and multilingual job postings to increase model generalizability. Enhancing the model with newer transformer architectures like DeBERTa or ELECTRA could yield better contextual understanding. A mobile application can be developed for broader accessibility. Incorporating explainable AI (XAI) techniques will help users understand prediction reasons, increasing trust. Continuous learning pipelines can be introduced to update the model as new fraud patterns emerge. Integration with major job portals like LinkedIn or Indeed can offer proactive fraud prevention. Improving the interface for better user experience and accessibility is also planned. IoT or browser-based input tracking can support fraud detection based on user interaction behavior. Security enhancements like CAPTCHA and multi-layered login can protect the platform. Finally, collaboration with HR platforms and cybersecurity agencies may help in real-world deployment and validation.

6.1 Expanded Gesture Dataset and Variability Handling

To improve model robustness, future work will involve collecting a larger and more diverse dataset that includes multilingual job postings, various industries, and global regions. Real-time job data scraping tools can be integrated to ensure continuous updates. Data augmentation techniques such as paraphrasing, synonym replacement, and random insertion or deletion will be applied to enrich the training data and introduce variability. These methods will help the model generalize better across different writing styles and fraud patterns. Moreover, collecting user feedback on

predictions can be used to refine the dataset through active learning. This will not only expand the data but also improve its quality and relevance over time..

6.2 Motor Control Optimization and Real-Time Execution

To enhance performance and reduce computational cost, future efforts will focus on optimizing the model architecture and training pipeline. Lightweight transformer models such as DistilBERT, TinyBERT, or MobileBERT will be explored to ensure faster inference without significant accuracy loss. Techniques like hyperparameter tuning, early stopping, and learning rate scheduling will be employed for efficient training. Additionally, model compression strategies such as pruning, quantization, and knowledge distillation will be applied to reduce model size while maintaining performance. These improvements will enable the deployment of fraud detection models on low-resource environments such as mobile devices or edge systems. Overall, the goal is to achieve real-time predictions with minimal resource consumption.

REFERENCES

1. P. Kaur, "E-recruitment: A conceptual study", Int. J. Appl. Res., vol. 1, no. 8, pp. 78-82, 2015.
2. C. S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh and G. Deepakkumar, "Fake job detection and analysis using machine learning and deep learning algorithms", Revista Gestão Inovação e Tecnologias, vol. 11, no. 2, pp. 642-650, Jun. 2021.
3. A. Raza, S. Ubaid, F. Younas and F. Akhtar, "Fake e job posting prediction based on advance machine learning approaches", Int. J. Res. Publication Rev., vol. 3, no. 2, pp. 689-695, Feb. 2022.
4. Online Fraud, Jun. 2022, [online] Available: <https://www.cyber.gov.au/acsc/report>.
5. J. Howington, Survey: More millennials than seniors victims of job scams, Flexjobs, CO, USA, Sep. 2015, [online] Available: <https://www.flexjobs.com/blog/post/survey-results-millennials-seniors-victims-job-scams>.
6. Report Cyber, Jun. 2022, [online] Available: <https://www.actionfraud.police.uk/>.
7. S. Vidros, C. Kolias, G. Kambourakis and L. Akoglu, "Automatic detection of online recruitment frauds: Characteristics methods and a public dataset", Future Internet, vol. 9, no. 1, pp. 6, Mar. 2017.
8. S. Dutta and S. K. Bandyopadhyay, "Fake job recruitment detection using machine learning approach", Int. J. Eng. Trends Technol., vol. 68, no. 4, pp. 48-53, Apr. 2020.
9. B. Alghamdi and F. Alharby, "An intelligent model for online recruitment fraud detection", J. Inf. Secur., vol. 10, no. 3, pp. 155-176, 2019. □
10. S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur and R. Mourya, "ORFDetector: Ensemble learning based online recruitment fraud detection", Proc. 12th Int. Conf. Contemp. Comput. (IC3), pp. 1-5, Aug. 2019. □
11. I. M. Nasser, A. H. Alzaanin and A. Y. Maghari, "Online recruitment fraud detection using ANN", Proc. Palestinian Int. Conf. Inf. Commun. Technol. (PICICT), pp. 13-17, Sep. 2021.
12. Vemula, V. R. (2025). Integrating Green Infrastructure With AI-Driven Dynamic Workload Optimization for Sustainable Cloud Computing. In Integrating Blue-Green Infrastructure Into Urban Development (pp. 423-442). IGI Global Scientific Publishing.
13. C. Lokku, "Classification of genuinity in job posting using machine learning", Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 12, pp. 1569-1575, Dec. 2021. □
14. O. Nindyati and I. G. Bagus Baskara Nugraha, "Detecting scam in online job vacancy using behavioral features extraction", Proc. Int. Conf. ICT Smart Soc. (ICISS), vol. 7, pp. 1-4, Nov. 2019.
15. S. Kotsiantis, D. Kanellopoulos and P. Pintelas, "Handling imbalanced datasets: A review", GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1, pp. 25-36, 2006. □
16. M. Tavallaei, N. Stakhanova and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods", IEEE Trans. Syst. Man Cybern. C Appl. Rev., vol. 40, no. 5, pp. 516-524, Sep. 2010.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details