International Journal of Advanced Engineering Application
Volume No.2 Issue No 4 April 2025
ISSN NO:3048-6807

Smart Detection of Fake Job Post Using Albert + Xgboost Intelligence

Mrs. Siriginedi Sai Brundavanam[1], Malleswari Gujjula[2], Raavi Samba Siva Reddy[3], Yarra Harika[4], Dhupati Srinivasu[5]
[1] Assistant Professor, [2,3,4,5] Students, Department of Computer Science and Engineering, Dhanekula Institute of Engineering and Technology, Ganguru, AP, India

Abstract
The increase of false job adverts across the internet has resulted in a spike in employment fraud which greatly endangers a job seeker's finances as well as their identity. Many people fail to identify which job listings are authentic and which ones are fake, making them an easy target for fraudsters. This project proposes a solution in the form of a Fake Job Post Detection System which applies ALBERT for the extraction of deep contextual features from the job descriptions and XGBoost for categorizing the posts into real or fake. It is trained on a Kaggle dataset containing labeled job postings with details such as job title, company name, and description. The system takes user-inputted job details and identifies patterns associated with fake job listings. ALBERT extracts linguistic features and XGBoost makes predictions based on the extracted features. To build trust Explainable AI (XAI) is incorporated wherein the user can comprehend the decision-making process. This system not only assists job seekers in spotting fraudulent job advertisements but also encourages a safer job-seeking experience.

Keywords: - Fake job post detection, ALBERT, XGBoost, Hybrid architecture, Deep learning, Natural language processing (NLP).

INTRODUCTION

As job searching proceeds to embrace the online platform, the level of scams seems to have increase. The online scamsters had exploited every nook and corner of the digital platforms to defraud the unsuspecting candidates. New graduates and underemployed have likely become targets for these criminals to exploit them for financial loss, identity theft, and exploitation. The unfortunate part is that these kinds of fraud job placements are posted on job portals, social media, and through email campaigning, making it difficult to distinguish genuine from fake job openings.

To fix this problem, we suggest a Fake Job Post Detection System that relies on Natural Language Processing and Machine Learning techniques. The model harnesses the power of ALBERT for deep contextual feature extraction, combined with XGBoost, a powerful ensemble learning algorithm for classification. The model training dataset is built from Kaggle datasets , labeled with features relating to job information, such as title, company, and job description. Decision-making will be subjected to SHAP for explainable-AI (XAI) processes.

Conventional fraud detection methodologies such as manual screenings and scams reported by users are proving to be unfit for accommodating the emerging volume rates of job postings. In fact, AIs offer scalable and efficient alternatives. Past attempts at utilizing naive Bayes, Decision Trees, and Random Forest for similar classifications were moderately successful but faced class imbalances and poor generalization. While use of Transformer-based models such as BERT or RoBERTa improved along the way of fraud detection, it was a computational-intensive process while in parallel lacking interpretability.

The paper contemplates an efficient alternative utilizing ALBERT's lightweight transformer architecture coupled with robust classification of XGBoost and Explainable AI (XAI) for interpretability of the model add to job seekers' protection against scams. Furthermore, real-time validation of fraud is included with web scraping, enhancing the ability of the detection.

LITERATURE SURVEY

Multiple studies are going on for detecting fraudulent jobs using AI and machine learning-based techniques. According to *IEEE Xplore (2023)*, SVMs, Decision Trees, and Logistic Regressions are a good fit for the detection of fake job postings **Reference[1].** Another article on job posting description analysis was published by *IEEE Xplore (2024)* where they used CNN and LSTM techniques to validate their accuracy in fraud detection **Reference [4].**

The hybrid model increases accuracy by combining deep learning methods with standard classification algorithms to improve the result. As per the research, augmentation of traditional classifiers by a CNN and LSTM deep learning model would fare better on fraud detection **Reference[6].** Fraud detection primarily depends on the features being engineered; it has improved the best model using NLP techniques including TF-IDF, word embeddings, and sentiment analysis **Reference [9].** According to the research presented, Ensemble methods like Random Forests and Boosting Algorithms are more efficient than the conventional algorithms in fraud detection **Reference[8].**

SPRINGER (2025) uses a synergic approach that combines different statistical models of machine learning to detect frauds **Reference[2].** These AI-powered webs serve the role of filtering jobs that participate in fraud. *IJSET (2023)* will present a model with AI that filters job postings to preemptively prevent fraudulent applications being filtered by the system before it gets into the organization **Reference[3].** *JETIR (2024)* is concerned about a multilayer perceptron that uses NLP-recommended text analysis to adopt high-utility detection of fraud **Reference[10].**

Further studies emphasized hybridization of different approaches. *IEEE Xpert(2024)* **Reference*[5]*** describes Fake Job Post Detection in Machine Learning, and declares that it was an effective solution. IJFMR(2024) explains different kinds of models in classification techniques and suggests machine learning as better fraud detection tool **Reference[7].**

PROPOSED SYSTEM

The Fake Job Posting Detection System discriminates between real or fake job postings based solely on job description information. ALBERT model extracts deep features while XGBoost offers fast classification. Various Explainable AI techniques are then adopted to bring transparency to model predictions and assist users in establishing why a job posting is considered real or fake.
The dataset contains 18 columns, including Job ID, Location, Requirements, Skills, Job Role, Salary, Job Description, and Agreement, and other relevant fields.. The input text undergoes preprocessing that includes text cleaning, tokenizing, and removal of stopwords. The cleaned text is fed into the ALBERT model, which generates contextual embeddings that capture deep semantic meanings. These embeddings are input to the XGBoost classifier for an efficient way of categorizing the job post.
Trust and usability are improved as XAI techniques explain the classification results, indicating to the user which terms were most pertinent in influencing a decision. Thus, the solution embodies the principles of not merely accurate but also interpretable and trustworthy identification of spurious job postings, protecting unwary candidates from losses.

Fig1: A Conceptual Illustration of Identifying Job Offers

SYSTEM ARCHITECTURE AND WORKFLOW

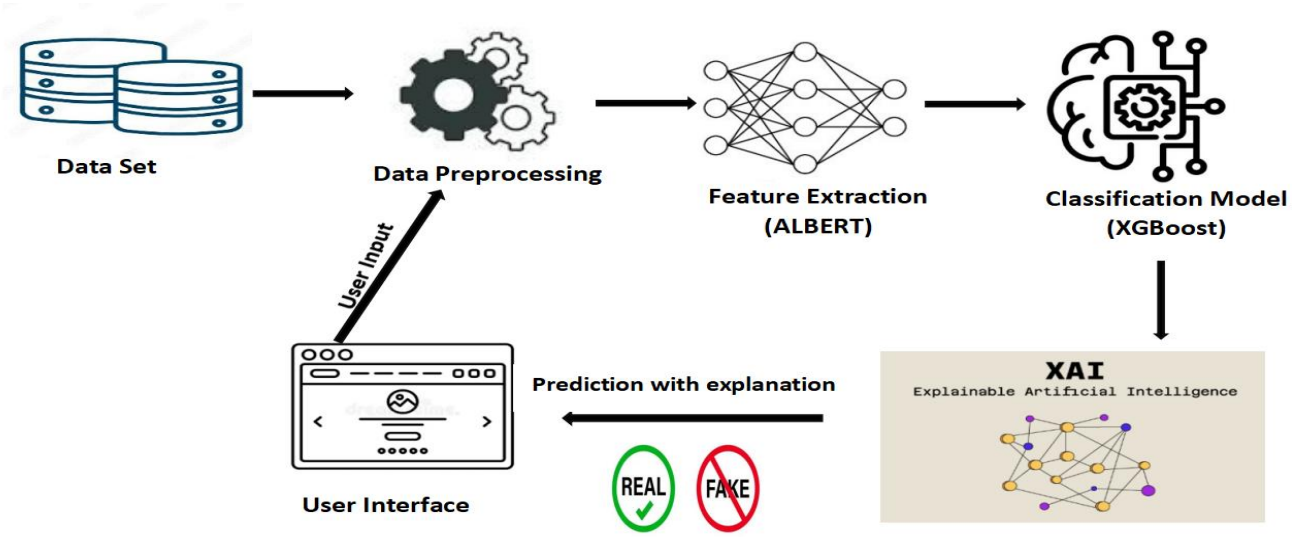

Fig2: System Architecture

SYSTEM WORKFLOW

1.Data Collection

The dataset used for training and evaluation purposes is collected from Kaggle's Fake Job Post Detection Dataset comprising labeled posts about fake jobs.

The dataset comprises two columns: Job description, which is text, and Class label-0 for real and 1 for fake. The input text undergoes preprocessing that includes text cleaning, tokenizing, and removal of stopwords.

2.Data Preprocessing:

Handles missing values using imputation techniques or deletion of all other incomplete records.

Also, application of NLP preprocessing techniques such as tokenization (split words from text), stopword removal (the irrelevant words are removed), stemming (the words are reduced to their root form), and lemmatization (the words are converted to their base form).

Text is standardized to lower case, removing distinctive characters for consistency.

3.Data Balancing with SMOTE:

SMOTE is the Synthetic Minority Over-sampling Technique for the balancing of the dataset.

Fake job postings are often found to be less, SMOTE will thus generate real synthetic samples, lest the model should show bias towards a real job posting.

By this treatment, the generalization of the model is improved, and misclassification probabilities due to class imbalance are also reduced.

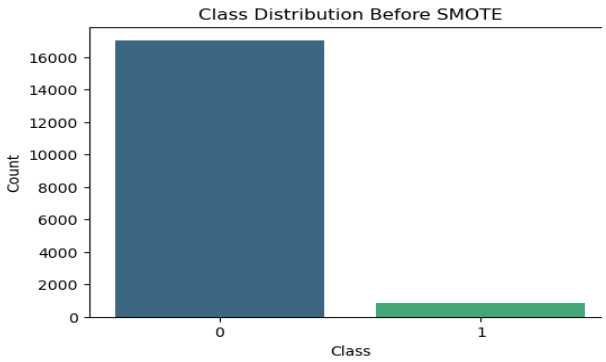

Fig 3: Class Distribution before SMOTE

Fig 3: Class Distribution after SMOTE

4.Feature Extraction using ALBERT:

ALBERT (A Lite BERT) is employed for creating deep contextual embedding from descriptions of jobs.
It encapsulates the semantic meaning of the text which helps in pinpointing slight disparities between genuine and fictitious job postings.
Pretrained weights for ALBERT have been fine-tuned on job text to give their best blend of being able to do fraudulent job postings with inaccurate characterizations.
Final classifier, the XGBoost classifier, takes the sample feature vectors for making the last decision.

5.Classification with the Use of XGBoost:

ALBERT-extracted embeddings are the training basis for the XGBoost (Extreme Gradient Boosting) classifier.
Structured and unstructured data format is the strong suit of XGBoost for fraud detection.
Gradient boosting inherent to XGBoost specializes in improving classification performance without increasing the risk for overfitting.
With the features extracted and metadata associated with the job post, the model predicts if that job post is real or fake.

6.Explanatory AI (SHAP):

SHAP (SHapley Additive Explanations) is one of the Explainable AI (XAI) techniques which makes use of prediction transparent.
SHAP Importance scores features helping in understanding how a job post would be classified as real or fake. Such feature encourages trust in the model as it provides understandable interpretations to predictions.

Model Evaluation

- **Accuracy(91.27%):** The model has rightfully classified 91.27% of the cases.
- **Precision- Real (100%) Fake (85%):** All real predicted cases are correct, but 85% of predicted fake cases are indeed fraudulent.
- **Recall- Real (83%) Fake (100%):** It identified 83% of real existing cases, and it identifies 100% of fraud cases.
- **F1-Score - Real Cases (90%), Fake Cases (92%):** It accounts for a balance of precision and recall whereby the overall
  robustness will be guaranteed.

```
2025-03-12 15:55:25,705 - INFO - Training XGBoost model...
2025-03-12 15:56:03,775 - INFO -  Model Accuracy: 91.27%
2025-03-12 15:56:03,807 - INFO -
              precision    recall  f1-score   support

           0       1.00      0.83      0.90      3403
           1       0.85      1.00      0.92      3403

    accuracy                           0.91      6806
   macro avg       0.92      0.91      0.91      6806
weighted avg       0.92      0.91      0.91      6806

2025-03-12 15:56:04,009 - INFO - XGBoost Model saved successfully
```

Fig 5: Classification report

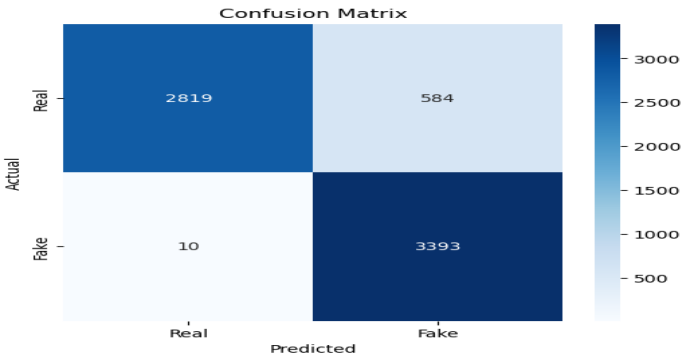

Fig 6: Confusion Matrix

User Interface

To access the system, users must create an account by registering and logging in. This ensures secure access and personalized tracking of analyzed job posts.
Once authenticated, users can submit job details in order to authenticate the job post.

Analysis & Feedback

The system processes the input by applying **text preprocessing** techniques such as cleaning, tokenization, and stopword removal. The **ALBERT model** extracts deep contextual embeddings, which are then classified using **XGBoost**.
Upon analysis, the system provides results:

- **Real or Fake Classification** – Job postings are flagged as either **real or fraudulent**.
- **SHAP-Based Explanations** – Users receive insights on why a job post was classified as fake.
- **Cautionary Advice** – If a post is detected as fraudulent, the system provides **warnings and tips** on avoiding scams.

This **secure, AI-powered** system helps job seekers **detect fraudulent listings, prevent scams, and make informed decisions** in their job search.

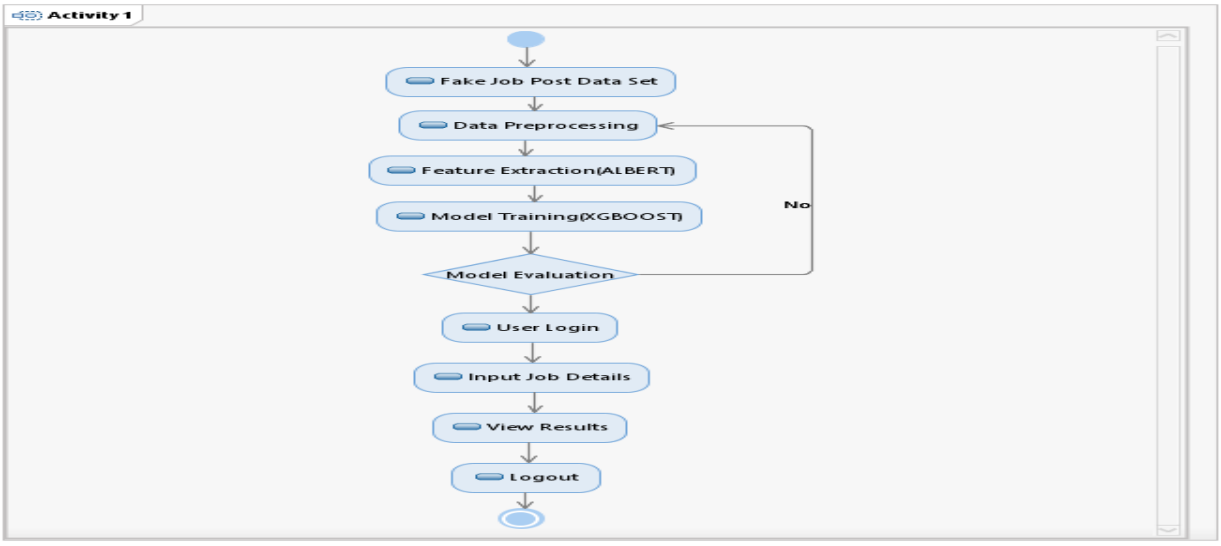

Fig7: Activity diagram

Advantages of the Proposed System:

Greater Accuracy: The deep contextual embeddings of ALBERT combined with the strong Classification capabilities of XGBoost allow for better detection of fake job posts.

Efficient: ALBERT reduces training time and expenditures in terms of computational resources, which makes the system efficient even with large data sets.

Explainability: With the use of XAI methods, such as SHAP, the system's predictions become more interpretable and accordingly more trustable.

User Interaction: Within this system, a job seeker has the capability of checking for authenticity of job postings and increasing the safety and confidence of online job recruitment.

Further Improvement of the System

The implementation of further intelligent features could include:

- Multi-language support
- Integration with recruitment platforms like LinkedIn , Naukri , Foundit.
- Feature engineering enhancement with ALBERT embeddings combined with metadata.
- Hyperparameter tuning and ensemble learning into XGBoost.
- A system of user reports for purposes of retraining and improving accuracy.

CONCLUSION

The Fake Job Post Detection System is a robust and viable system for combating fraudulent job postings. Its components of ALBERT for deep contextual feature extraction, XGBoost for classification, and Explainable AI (XAI) for interpretability of the model add to job seekers' protection against scams.

With a pleasant accuracy of **91.27%**, the model is highly effective in distinguishing between bona fide and fraudulent job postings.The improvement of such a system would keep it at pace with the constantly changing online recruitment environment and, thus, able to detect emerging fraud patterns.

REFERENCES

[1]. IEEE Xplore, ”Detection of Fake Online Recruitment Using Machine Learning Techniques”, https://ieeexplore.ieee.org/document/10074276

[2]. SPRINGER, “A machine learning approach to detecting fraudulent job types”, https://link.springer.com/article/10.1007/s00146-022-01469-0?utm_source=chatgpt.com

[3]. IJSET, “Fake Job Post Detection Website”, https://www.ijset.in/wp-content/uploads/IJSET_V12_issue5_799.pdf

[4]. IEEE Xplore, ”Online Recruitment Fraud (ORF) Detection Using Deep Learning Approaches”, https://ieeexplore.ieee.org/document/10614582

[5]. IEEE Xpert, ”Fake Job Post Detection using Machine Learning”, https://www.ieeexpert.com/python-projects/fake-job-post-detection-using-machine-learning/

[6]. ResearchGate, “ Fake Job Detection & Analysis Using Machine Learning & Deep Learning Algorithms”, https://www.researchgate.net/publication/352159024_Fake_Job_Detection_and_Analysis_Using_Machine_Learning_and_Deep_Learning_Algorithms

[7]. IJFMR, ”Fake Job Post Detection using Machine Learning”, https://www.ijfmr.com/papers/2024/1/13906.pdf

[8]. ACM Digital Library, “Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches”,https://dl.acm.org/doi/10.1007/s11063-021-10727-z\

[9]. ResearchGate, “A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques”, https://www.researchgate.net/publication/349884280_A_Comparative_Study_on_Fake_Job_Post_Prediction_Using_Different_Data_mining_Techniques

[10]. JETIR, “Prediction of Fake Job Ad using NLP-based Multilayer Perceptron”, https://www.jetir.org/papers/JETIR2404340.pdf

[11]. SPRINGER, “Fraud-BERT: transformer based context aware online recruitment fraud detection”, https://link.springer.com/article/10.1007/s10791-025-09502