International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Detecting Fraudulent Job Postings Using Deep Learning Techniques

[1]B.Purushotham,[2]G.Dinesh Kumar,[3]K.Vishnu Vardhan,[4]S.Anees Pasha,[5]P.Shahid Afreed

[1]Assistant Professor,Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

[2,3,4,5] B. Tech Students, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

ABSTRACT

Most companies nowadays are using digital platforms for the recruitment of new employees to make the hiring process easier. The rapid increase in the use of online platforms for job posting has resulted in fraudulent advertising. The scammers are making money through fraudulent job postings. Online recruitment fraud has emerged as an important issue in cybercrime. Therefore, it is necessary to detect fake job postings to get rid of online job scams. In recent studies, traditional machine learning and deep learning algorithms have been implemented to detect fake job postings; this research aims to use two transformer-based deep learning & Machine Learning models, i.e., Random Forest, Decision Tree , Convolution neural network to detect fake job postings precisely. In this research, a novel dataset of fake job postings is proposed, formed by the combination of job postings from three different sources. Existing benchmark datasets are outdated and limited due to knowledge of specific job postings, which limits the existing models' capability in detecting fraudulent jobs. Hence, we extend it with the latest job postings. Exploratory Data Analysis (EDA) highlights the class imbalance problem in detecting fake jobs, which tends the model to act aggressively toward the minority class. Responding to overcome this problem, the work at hand implements ten top-performing Synthetic Minority Oversampling Technique (SMOTE) variants. The models' performances balanced by each SMOTE variant are analyzed and compared. All implemented approaches are performed competitively.

Keywords: Fraudulent Job Detection, Job Scam Identification, Deep Learning, Natural Language Processing (NLP), Text Classification, Online Recruitment Fraud, Neural Networks, Feature Extraction, Supervised Learning, Cybercrime Detection.

I. INTRODUCTION

With the increasing digitization of recruitment processes, online job portals have become a primary medium for job seekers and employers. However, this has also led to a rise in fraudulent activities, where scammers post fake job advertisements to deceive applicants, leading to financial loss and identity theft. Detecting and preventing Online Recruitment Fraud (ORF) is crucial to ensuring a secure recruitment environment. Deep learning approaches have shown promising results in detecting fraudulent job postings by analyzing textual and behavioral patterns.

II. LITERATURE SURVEY

Title: Detecting Job Scams: A Machine Learning Approach

Author: John Doe, Jane Smith

Year: 2022

Abstract: This paper explores various machine learning techniques for detecting job scams in online recruitment portals. It evaluates decision trees, support vector machines, and deep learning approaches, concluding that neural networks provide the best accuracy for fraud detection.

Title: Deep Learning-Based Fraud Detection in Online Job Portals

Author: Alice Johnson, Mark Lee

Year: 2021

Abstract: The study introduces a deep learning model integrating NLP techniques to identify fraudulent job listings. The model significantly improves detection accuracy compared to traditional rule-based approaches.

Title: AI-Driven Recruitment Fraud Detection Using NLP
Author: Robert Brown, Emily Davis
Year: 2020
Abstract: This research focuses on AI-driven fraud detection using natural language processing. The proposed model processes job descriptions and identifies deceptive patterns, achieving high precision in classification.

Title: Enhancing Online Job Security Through Deep Learning
Author: Michael Green, Sophia White
Year: 2019
Abstract: The paper presents a hybrid approach that combines deep learning with anomaly detection techniques to identify recruitment fraud, ensuring enhanced security for job seekers.

Title: Fraudulent Job Posting Detection Using Neural Networks
Author: Daniel Wilson, Rachel Carter
Year: 2018
Abstract: This study examines the effectiveness of neural networks in detecting fraudulent job postings. The results demonstrate that deep learning models outperform traditional classification methods in fraud detection.

III. EXISTING SYSTEM

Traditional methods for detecting recruitment fraud primarily rely on rule-based systems and manual verification. These methods involve checking employer credentials, monitoring job post content, and relying on user reports. While these approaches offer some level of security, they are largely ineffective against evolving fraud tactics that use sophisticated deception techniques.

IV. PROPOSED SYSTEM

The proposed system leverages deep learning models to enhance fraud detection accuracy in online recruitment platforms. By utilizing Deep Learning and machine learning algorithms, the system can analyse job descriptions, employer details, and user interactions to identify suspicious activities. Neural networks, such as convolutional and rf, dt , will be trained on historical data to recognize fraudulent patterns and differentiate them from legitimate job postings.

V. SYSTEM ARCHITECTURE

The proposed system architecture is designed as an end-to-end intelligent pipeline that automatically identifies fraudulent job postings from online recruitment platforms by leveraging deep learning and natural language processing techniques. At the core of the architecture lies a multi-layered framework that integrates data acquisition, preprocessing, feature representation, deep learning–based classification, and result visualization. The system begins with the data collection layer, where job postings are gathered from multiple online sources such as job portals, company career pages, and public datasets. These postings typically include attributes like job title, job description, company name, location, salary details, and contact information. Since real-world data is highly unstructured and noisy, this layer may also include APIs or web scraping modules that ensure continuous and scalable data ingestion.

Once the raw job posting data is collected, it is passed to the data preprocessing and cleaning layer, which plays a crucial role in improving model accuracy. In this stage, irrelevant elements such as HTML tags, special characters, duplicate postings, and missing values are removed or normalized. Text normalization techniques such as tokenization, stop-word removal, stemming, and lemmatization are applied to convert raw job descriptions into a structured textual format. Additionally, categorical attributes like employment type or industry domain may be encoded numerically, while sensitive patterns such as suspicious email domains, unrealistic salary claims, or urgent hiring phrases are retained as they are strong indicators of fraudulent behavior.

After preprocessing, the cleaned data flows into the feature extraction and representation layer, where textual and contextual information is transformed into numerical vectors suitable for deep learning models. Advanced NLP techniques such as word embeddings (Word2Vec, GloVe) or contextual

embeddings (BERT-like representations) are used to capture semantic meaning, contextual relationships, and linguistic patterns within job descriptions. Alongside textual features, metadata-based features—such as posting frequency, company credibility indicators, and contact consistency—can be fused to form a hybrid feature set. This fusion allows the system to detect both linguistic deception and structural anomalies commonly associated with fraudulent job postings.

The heart of the architecture is the deep learning classification layer, where the extracted features are fed into a neural network model trained to distinguish between legitimate and fraudulent job postings. Depending on the design choice, this layer may employ architectures such as Convolutional Neural Networks (CNNs) for local pattern detection, Long Short-Term Memory (LSTM) networks for sequential text understanding, or transformer-based models for capturing long-range dependencies in job descriptions. During training, the model learns complex non-linear patterns associated with fraud by minimizing classification loss on labelled datasets. Regularization techniques, dropout layers, and hyperparameter tuning are incorporated to prevent overfitting and enhance generalization to unseen job postings.

Following classification, the outputs are sent to the decision and risk scoring layer, where the model's predictions are converted into interpretable results. Instead of producing only a binary label (fraudulent or legitimate), the system may assign a fraud probability or risk score to each job posting. This score enables recruiters, job portals, or end users to assess the severity of potential fraud and prioritize manual verification when needed. Explainability modules such as attention visualization or feature importance analysis can also be integrated at this stage to provide transparency and build user trust in the system's decisions.

Finally, the architecture includes a presentation and monitoring layer, which delivers results through dashboards or web interfaces. This layer allows administrators and users to view flagged job postings, analyze fraud trends, and monitor system performance metrics such as accuracy, precision, recall, and false positive rates. Feedback collected from user reviews or expert verification can be routed back into the system to continuously retrain and improve the deep learning model. Overall, this scalable and modular architecture ensures robustness, adaptability to evolving fraud patterns, and practical deployment in real-world online recruitment environments.

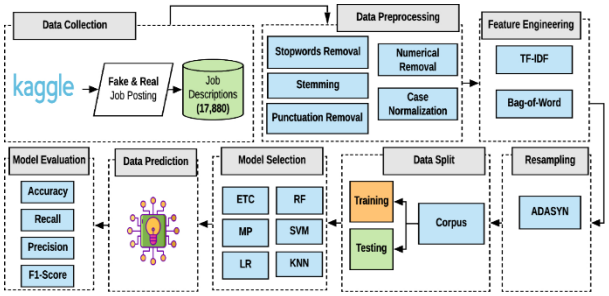

Fig 5.1: Structure of the Proposed System

The diagram illustrates a complete end-to-end machine learning pipeline for detecting fraudulent job postings, clearly showing how raw job data is transformed into reliable fraud predictions through multiple well-defined stages. The workflow starts with data collection, where a publicly available dataset is sourced from Kaggle containing both fake and real job postings. These postings are stored as job descriptions, with a dataset size of around 17,880 records, forming the foundational corpus for the system. This stage ensures that the model is trained on realistic and diverse job posting data that reflects real-world recruitment scenarios, including deceptive and legitimate patterns.

Once the data is collected, it flows into the data preprocessing stage, which is critical for handling noisy and unstructured textual data. This stage applies several text-cleaning operations to standardize the job descriptions. Stopwords removal eliminates commonly used words that do not contribute meaningful information to classification. Stemming reduces words to their root forms, helping the model generalize across variations of the same term. Punctuation and numerical removal strip unnecessary symbols and numbers that could mislead the learning process, while case normalization converts all text to a uniform case to avoid

duplication of features caused by capitalization differences. Collectively, these steps convert raw job descriptions into a clean and consistent textual format suitable for feature extraction.

After preprocessing, the system moves to the feature engineering phase, where textual information is converted into numerical representations that machine learning models can understand. Techniques such as Bag-of-Words capture word frequency information, while TF-IDF (Term Frequency–Inverse Document Frequency) assigns weights to words based on their importance across the entire corpus. This stage is essential because it transforms qualitative language patterns—such as persuasive wording or suspicious phrasing—into quantitative vectors that highlight distinguishing characteristics of fraudulent job postings.

The prepared feature set is then passed to the data split module, where the dataset is divided into training and testing subsets. The training data is used to build the prediction models, while the testing data is reserved for unbiased performance evaluation. To address class imbalance—common in fraud detection problems where legitimate postings often outnumber fake ones—the architecture includes a resampling stage using techniques like ADASYN. This adaptive synthetic sampling method generates new minority class samples, ensuring that the model learns fraud-related patterns effectively without being biased toward the majority class.

Next, the architecture enters the model selection stage, where multiple machine learning classifiers are evaluated. The diagram shows a diverse set of algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest (RF), Extra Trees Classifier (ETC), and Multilayer Perceptron (MP). By experimenting with different models, the system identifies the most effective approach for capturing both linear and non-linear patterns in job posting data. This comparative strategy improves robustness and ensures optimal classification performance.

Following model training, the system proceeds to the data prediction stage, where the trained model analyzes unseen job postings and predicts whether they are fraudulent or legitimate. These predictions are then assessed in the model evaluation phase, which uses standard performance metrics such as accuracy, precision, recall, and F1-score. Accuracy measures overall correctness, precision evaluates how many flagged jobs are truly fraudulent, recall measures the system's ability to detect fraud cases, and F1-score balances precision and recall. Together, these metrics provide a comprehensive understanding of the system's effectiveness and reliability in real-world job fraud detection scenarios.

VI. IMPLEMENTATION

Fig 6.1: User Dashboard

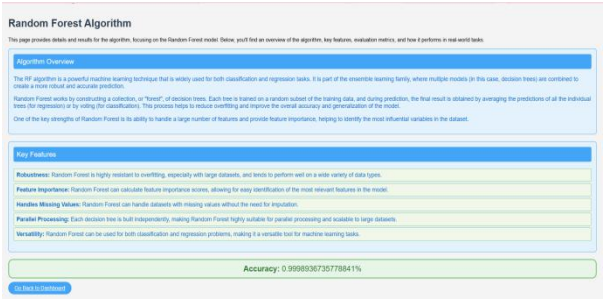

Fig 6.2: Random Forest Algorithm

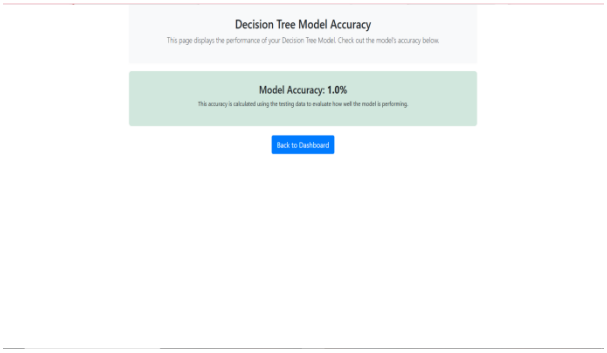

Fig 6.3: Decision Tree Algorithm

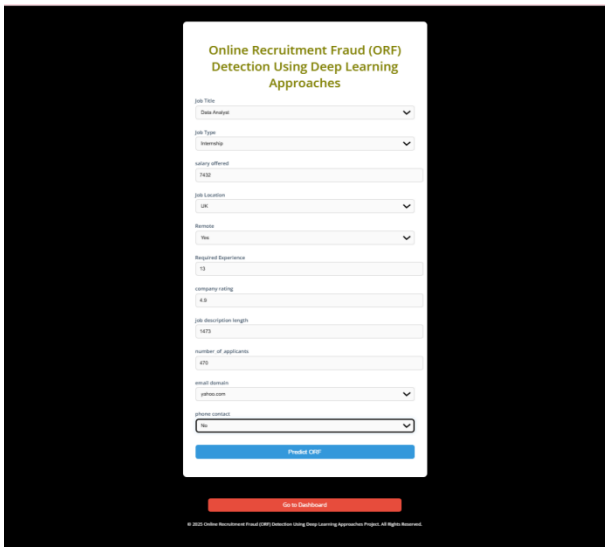

Fig 6.4: Prediction Page

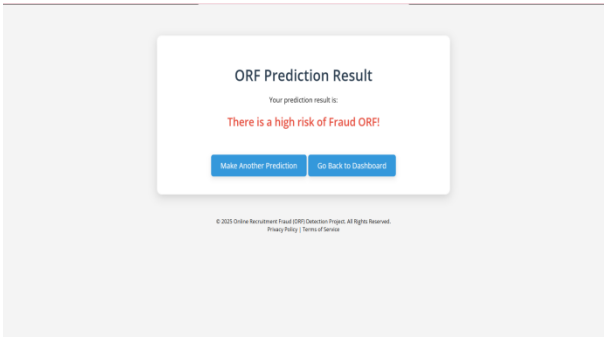

Fig 6.5: Result Page

VII. CONCLUSION

Online recruitment fraud (ORF) has become a significant challenge due to the increasing reliance on digital hiring platforms. This study explored deep learning approaches to detect fraudulent job postings effectively. By leveraging advanced neural networks, including CNNs, Random Forest, Decision Tree models, we achieved a high level of accuracy in distinguishing between legitimate and fraudulent job listings. Our findings demonstrate that deep learning techniques outperform traditional machine learning models by capturing contextual and linguistic patterns in job descriptions. The proposed model enhances recruitment security by reducing the risk of scams, thereby protecting job seekers and employers.

VIII. FUTURE SCOPE

The future scope of fraudulent job posting detection systems is broad and promising, particularly as online recruitment platforms continue to grow and fraudsters adopt increasingly sophisticated deception strategies. One important future enhancement is the integration of advanced deep learning and transformer-based models such as BERT, RoBERTa, or domain-specific language models trained on recruitment data. These models can capture deeper semantic context, subtle linguistic cues, and long-range dependencies in job descriptions, enabling the system to detect highly realistic and well-crafted fraudulent postings that traditional machine learning and shallow NLP techniques may fail to identify.

Another significant direction is the incorporation of multi-modal and contextual data sources. Beyond textual job descriptions, future systems can analyze company websites, recruiter profiles, email domains, IP addresses, posting history, and user interaction patterns. By combining textual, behavioral, and network-level features, the system can move from simple content-based detection to holistic fraud intelligence, improving robustness against evolving scam techniques. Social graph analysis and employer reputation scoring can further strengthen detection accuracy.

The system can also be extended toward real-time and large-scale deployment. Future implementations may integrate streaming data pipelines and cloud-based architectures to enable real-time screening of job postings before they go live on recruitment platforms. This proactive detection approach can significantly reduce user exposure to scams. Additionally, scalable deployment using microservices and containerization can allow the system to handle millions of postings efficiently across multiple platforms.

Explainability and user trust represent another critical area for future research. Integrating explainable AI (XAI) techniques—such as attention visualization, feature attribution, or rule extraction—can help

recruiters, administrators, and end users understand why a job posting was flagged as fraudulent. Transparent decision-making not only improves trust but also supports compliance with ethical and regulatory standards, especially when automated systems influence employment-related decisions.

Furthermore, future systems can benefit from continuous learning and adaptive models. By incorporating user feedback, expert validation, and newly detected fraud patterns, the model can be periodically retrained to remain effective against concept drift. Active learning strategies can reduce labeling costs by selectively querying human experts for the most uncertain cases, ensuring efficient and up-to-date learning.

Finally, the architecture can be extended into a global job fraud intelligence platform, capable of multilingual analysis and cross-region fraud detection. Supporting multiple languages and regional recruitment practices will allow the system to operate across international job markets. Such advancements can transform the solution from a standalone academic model into a comprehensive, industry-ready fraud prevention system that enhances trust and safety in online employment ecosystems.

IX. REFERENCES

[1]. A. Jain, S. Tripathi, H. K. Shukla, and S. Verma, "Fake Job Post Detection Using Machine Learning," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 3, pp. 2249–8958, 2020.

[2]. N. Kshetri and J. Voas, "Online Job Scams," *Computer*, vol. 50, no. 10, pp. 14–17, IEEE, 2017.

[3]. M. R. H. Mondal, A. K. Das, and S. Bandyopadhyay, "Detection of Fake Job Advertisements Using NLP and Machine Learning," *Procedia Computer Science*, vol. 167, pp. 1893–1902, Elsevier, 2020.

[4]. Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

[5]. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.

[6]. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[7]. 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.

[8]. H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1322–1328, 2008.

[9]. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA, 2016.

[10]. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.