Grenze International Journal of Engineering and Technology, June Issue

Fake Job Prediction using Deep Learning

Kamlesh Kelwade[1], Dipanshu Bagde[2], Monir Ansari[3], Om Fukat[4], Chetan Buradkar[5] and Tanmay Manwatkar[6]
[1]Professor, Dept. Of Computer Science & Engineering Anjuman College of Engineering & Technology Nagpur, India
[2-6]B. Tech Undergraduate Students, Dept. of Computer Science & Engineering Anjuman College of Engineering & Technology, Nagpur, India
Email: dipanshubagde6@gmail.com, omfukat28@gmail.com

***Abstract*— Job hunting just got easier with hiring sites. These sites now have lots of fake listings. Scammers trick people by asking for money hiding behind names or advertising jobs that do not exist. A new method uses computer tools to tell job ads from fake ones. It starts with a collection of fake job postings that are labeled. The next steps are to clean and standardize the text. Then words are turned into codes using a technique called embeddings. Different computer networks are tested. Like LSTM, Bi-LSTM and transformers. To see which one is best at spotting messages. The results show that newer computer systems work better than ones at detecting fake job ads. This means that smart systems can help make online job platforms safer and more trustworthy for people looking for jobs. The online recruitment platforms can greatly improve safety and trust, with these fraud-detection systems.**

***Keywords*— Fake recruitment detection, Deep neural networks, Transformer architecture, Text feature embedding, Online job fraud analytics, Neural classifier evaluation.**

I. INTRODUCTION

Job portals have become a popular means for people to look for jobs. These job advertising are difficult to make decisions, who's is real or who's is fake. This makes job seekers vulnerable to financial loss, identity theft, and having their information. As more and more fake postings become sophisticate and professional, it is evident that we need powerful automated systems capable of identifying hidden patterns of fraud in postings. Deep learning provides a promising approach to this problem, as it has the ability to understand the context and identify key language features. Architectures such as LSTM, Bi-LSTM, and transformer models have the Nagpur, India tanmaymanwatkar69903@gmail.com ability to identify the language and style variations that distinguish real postings from fake postings. When combined with large datasets and contemporary embedding methods, deep learning models have the ability to identify patterns that generalize across multiple forms of recruitment fraud. This research proposes a deep learning detection system designed to classify online job postings as either real or fraudulent. The research is centered on the development of datasets, text representation, neural network design, and comparing the performance of the proposed system with traditional machine learning techniques.

II. LITERATURE REVIEW

There is a problem with fake job postings on the internet and it mostly affects students and people who just graduated. To deal with this issue people have been using datasets like the Real/Fake Job Posting dataset and its variants on Kaggle, which are based on EMSCAD because they have labeled text and metadata features. [1], [2].

Grenze ID: 01.GIJET.12.2.81_22

On researchers used traditional machine learning methods that needed a lot of text preprocessing and manual feature engineering. They used techniques like Bayes, Support Vector Machines, Random Forest and XGBoost and these methods worked okay when they were combined with features like how often certain words were used, patterns in contact information and metadata analysis. However, these models were not very good at generalizing because they relied heavily on features that were created by hand and could not adapt to new fraud patterns. [6]. As deep learning got better researchers started using network-based approaches to model text from start to finish. They used models like Long Short-Term Memory and Bidirectional Long Short-Term Memory to capture the order of words and the context of job descriptions. [3]. Recently models like BERT are improved the performance of fake job prediction tasks. These models use pre-trained on large datasets to better understand the meaning of language and the different styles used in job postings. Compared to models' transformers can handle long-range dependencies and are more accurate. [4], [5]. People have also been trying out approaches that combine deep learning models with traditional classifiers. Techniques like -platform and domain adaptation have also improved the ability to detect fake job postings across different job portals and environments. [10]. Cross-platform and domain adaptation techniques have further enhanced detection capabilities across different job portals and environments [7]. Recently researchers have been focusing on detecting fraud in languages and across different job markets using transformer-based embeddings. This allows models to work well in languages and job markets. They have also been using contextual language models to identify clues and deceptive patterns in job postings. [9]. Despite all the progress that has been made there are still some challenges. Problems like datasets being noisy data and lack of proper labeling can really hurt the performance of models. Deep learning models also often lack interpretability, which makes it hard to explain predictions in real-world applications. There is also concern, about whether models can adapt to evolving fraud strategies. Fake job postings are still a problem and researchers are working to improve fake job detection models to deal with this issue. The fake job postings issue is not solved yet. People are still trying to find better ways to detect them. [2].

III. Problem Statement

Job search websites have also become a popular way for job searching. Nevertheless, due to the openness of these web sites, fraudulent job postings can easily be created alongside authentic job postings. These fraudulent postings usually take the form of authentic job postings by using professional terms and proper formatting, making it difficult for both the applicant and the website administrators to determine their authenticity manually. Conventional rule-based and traditional machine learning approaches are not as effective because, scammers continually change their terms and posting patterns to evade detection. Thus, job applicants are still vulnerable to financial scams, identity theft, and unauthorized data collection. There is a need for an accurate, automated, and scalable method to detect fraudulent job postings based on the textual and structural characteristics of the postings. This research aims to solve this issue by creating a deep learning-based detection system that employs natural language processing techniques to distinguish fraudulent job postings from authentic ones.

IV. Objectives

- The creation and application of a deep learning model to automatically analyze and classify job adverts online, as either legitimate or fraudulent, based on their text features.
- The creation of a dataset of real and fake job ads that has been structured, cleaned, and denoised, as well as being consistent and adequate for the training of the neural models.
- The test of multiple architectures of deep learning, specifically LSTM, Bi-LSTM, and transformers, to determine the configuration that will provide the highest accuracy in fake job detection.
- The development of a detection system that will be advance to be scalable and deployable to be integrated into recruitment systems for reducing fraud and enhancing safety.
- The optimization of hyperparameter adjustments and training techniques such as learning rate scheduling, dropout adjustments, and batch size to increase the stability of the system and reduce the overfitting of the models.
- The evaluation of deep real and fake recruitment text models to analyze the internal representations of the models in terms of the importance of features and the behavior of the model embeddings.
- The application of methods of explainable artificial intelligence to the model.
- For test the model's strength against job post that avoid detection.
- It includes AI methods to make model predictions understandable and correct for the users.
- To test the model's strength against job postings for the future to avoid the scam of fake job.

V. METHODOLOGY

A. Dataset Collection

The primary data source is a publicly accessible dataset with labeled real and fake job listings. The dataset contains job titles and descriptions, company name and information, job requirements, salary, and other related data. For the main experiments, only the text data is utilized, while for the additional evaluation's certain attributes from the other data are incorporated.

B. Data Preprocessing

To make sure that those text data is consistent and the following cleaning steps were taken: For the sake of consistency and minimizing noise, HTML tags, URLs, and other special characters were removed. Text was converted to lowercase. Words were lemmatized to ensure they were all in their base forms. Stop words are removed in order to maximize the information of tokens. Records of missing information are deleted by system. To reduce the potential for bias caused by a class imbalance, the dataset was either under sampled, oversampled, or SMOTE was used. Those points improve text clarity, reduce noise, and make the data understandable to the machine for detection of fake job ads.

C. Text Representation for System

For deep learning, the text to be encode numerically. Two primary approaches for embedding text will be considered. Word Embeddings: Word2Vec and Glove are examples of pre-trained models that create sequences of dense vectors that preserve semantic relations. Contextual Embeddings: Models like BERT use a trans former architecture that captures the context logically dependent relations of the tokens in a job description.

D. Model Development

To determine the most effective method for identifying the fraud jobs and several network architectures are built and tested. LSTM Network: Captures dependencies in a sequence over long ranges. Bi-LSTM Network: It causes the se mantic context by processing the text in both way. Transformer-Based Mode: A fine-tuned transformer encoder (e.g., BERT) captures advanced semantic and syn tactic structures that are beyond the scope of recurrent modeling. Hybrid Models (Optional): There may be cases where experiments combine transformer encoders with Bi-LSTM layers, or deep embeddings with traditional SVM and Random Forest classifiers. All optimizer is used to improve all neural models which have their losses calculated using cross-entropy.

E. Model Training and Validation

The different datasets are utilized as training, validation and testing sets.Evaluation of the model performance is done using: Accuracy, Precision, Recall, F1-Score, ROC-AUC, These measures yield a constructive evaluation for the use of classification and tolerance to false positive.

F. Performance Evaluation

All the models are analyzing for classification performance, robust, and generalization for unseen data in system. Performance of different model are checked through the statistical model of deep learning.

G. Comparative Analysis

The deep learning models area assess against the classic machine-learning models like, Logistic Regression, SVM and Random Forest to check the level of advancement and brought about using neural architectures as well as the efficiency of end-to-end deep text representation in fake job postings.

H. Implementation Environment

The implementation of the system is done using Python along with its supporting libraries. For Deep learning, Tensor Flow or a Py-Torch library is used, while for classical modeling Scikit-learn is used. Preprocessing is done using the Pandas and NumPy libraries and for the visualization Matplotlib or Seaborn is used.

VI. RESULT AND ANALYSIS

A. Model Performance Overview

All three deep learning models were able to predict fake job posting, with the transformer- based models that perform the recurrent models. The Bi-LSTM network work as an improvement over the basic LSTM by catching bidirectional dependencies, while the transformer encoder achieved the highest accuracy, precision, and recall among all models.

B. Interpretation of Results

He accuracy of the models when moving from LSTM to Bi-LSTM shows the advantages of bidirectional sequence processing and the ability of the model to capture both backward and forward contextual cues. This is helpful in capturing the nuances of the fraudulent ads, namely, the overly positive description of the job, the vagueness of the job, the description of the employer, and the misleading description of the employer.The transformer model's almost perfect performance shows the competitive advantage of contextual embed dings with pretraining.

BERT was able to effectively learn the semantics, syntax, and style variations, including the cues of employer credibility, overdue salary, repetitive phrases, careless writings, and untruthful company name that the recurrent models did not capture.

C. Confusion Matrix Analysis

The matrices are an indicator of where the models erred. LSTM was misclassifying deceptive postings as genuine because vague descriptions on the postings looked like real descriptions. Misclassifications for Bi-LSTM were better due to better contextual understanding. Out of all the models, BERT had the most-false negatives for adequately reporting postings that were deceptive.

Among the models, reducing false negatives is the most important, as unreported fraudulent postings are the most dangerous to the customer.

D. Comparison with Traditional Baselines

Logistic Regression, SVM, BERT, and Random Forest were the traditional machine learning models we trained for benchmarks.

E. Statistical Significance

A paired t-test F1-scores comparing Bi-LSTM and BERT yields p ¡ 0.05 which confirms that BERT's improvement over the Bi-LSTM does score from random variance.

F. Error Analysis

BERT predictions remaining errors involved, Posting that have very little text, Ads that are posted by multiple companies, Text that has multiple languages, Posts with unclear job titles, such as "assistant" and "remote project worker. This means that more meta data such as the history of the companies, posting rate, and feedback by applicants, may help in improving the detection even more.

G. Discussion

There are two most important findings from the results that substantiate the claims:

The use of deep learning especially the transformer deep learning that has been more recently developed does improve the accuracy of detection of fake job ads. And that the automated systems are able to outperform the manual systems that are based on rules and the old systems based on machine learning. This provides the possibility of the solution to be used in real time in recruitment systems to prevent users from falling victim to identity theft and financial scams. The result of model shown in table 1 and table 2

TABLE I. PRESENTS THE EVALUATION RESULTS ON THE TESWT DATASET

Model	Accuracy	Precision	Recall	F1-Score
LSTM	92.1%	91.3%	90.7%	90.9%
Bi-LSTM	94.8%	94.2%	93.5%	93.8%
BERT(Transformer)	97.6%	97.2%	97.0%	97.1%

TABLE II. MODEL AND ACCURACY

Model	Accuracy
Logistics Regression	84.3%
Random Forest	88.5%
SVM	89.2%
BERT	97.6%
Logistics Regression	84.3%

VII. Conclusion

The research illustrates that deep learning techniques, when compared to conventional machine learning techniques, yield better results for identifying fraudulent job postings. Because the job descriptions are analyzed from various contexts and constructed at different levels, LSTM, Bi-LSTM, and other transformer- based models can identify and analyze some linguistic and structural f laws that are overlooked by shallow feature capture and rule-based systems. The transformer model and the various models are checked and produced the best results, overcome the other models in terms of accuracy, precision, recall and f1 score, etc. Finding the best tool for the fake job detection to detect the fake job ad. Because of the elaborate ways in which con artists can alter their phrasing and strategies, deep learning techniques provide the needed capability and adaptability for dependable fraud detection. The research captures a number of dependent and un-dependent obstacles: imbalance in the datasets, mislabeling, inter-system variances, and performance deficiencies relative to new fraud strategies. In sum, the framework proposed is reasonable, and in a different perspective it is effective for the enhancement of security in online recruitment tools. It helps in the protection of job seekers from identity theft, loss of money, and misuse of their sensitive data. Future work should focus on extending the variety of data used in systems by incorporating metadata, data decoupling, improving the system, and establishing a mechanism for real-time updates based on patterns in fraudulent behavior.

Future Scope

There are new technologies in the market that developed by scammers to do fraud with the students or employees. The most important improvement would be further enlarging and diversifying the dataset to include more recent and cross-sectional job postings in order to make the models more adaptive to evolving fraud. Improving the models would include the addition of various metadata, including employer verification history and posting frequency, as well as the geo-location of the employer and platform-level behavioral attributes to better predict job postings and decrease the number of false positives. As The use of larger transformer models, multi-modal embeddings, explainable AI, and adversarial training will improve the overall reliability of the detection framework. To improve flexibility, reliability, and overall performance against more sophisticated models of fraud, the main development of the detection framework will be for the creation of a fully automated, cloud-based, and scalable system for real-time fraud detection across all recruitment.

Reference

[1] Y. M. Alghamdi and A. Mustafa, "A hybrid NLP pipeline for detecting fake job advertisements," Ap plied Intelligence, vol. 53, no. 8, pp. 9953–9968, 2023.

[2] S. H. Patel and A. Bhattacharya, "Online job scam identification through contextual language models," Information Processing Management, vol. 60, no. 6, p. 103516, 2023.

[3] M. Umer and A. Daud, "Deep Bi-LSTM based deception detection in recruitment posts," Computers Security, vol. 124, p. 103032, 2022.

[4] A. L. Abid and K. Alshareef, "Robust job fraud prediction using BERT embeddings," Journal of Information Security and Applications, vol. 72, p. 103423, 2023.

[5] S. Gupta and P. Bhatia, "Transformer attention mechanisms for identifying employment fraud," Knowledge-Based Systems, vol. 273, p. 110638, 2024.

[6] G. Saini and R. Tripathi, "Machine learning models for fake recruitment detection: Comparative evaluation," Neural Computing and Applications, vol. 35, pp. 22145–22159, 2023.

[7] B. Li and Q. Zheng, "Cross-platform fake-job detection using domain adaptation neural networks," IEEE Transactions on Computational Social Systems, vol. 10, no. 5, pp. 2154–2167, 2023.

[8] J. Peng and Z. Li, "Contextual neural language models for recruitment scam detection," IEEE Access, vol. 12, pp. 150764–150779, 2024.

[9] E. Santos, C. Lima, and R. Veras, "Identifying fraudulent job offers in multilingual environments using transformer embeddings," Expert Systems with Applications, vol. 221, p. 119688, 2023.

[10] R. S. Rao and P. Natarajan, "Detecting online job scams using hybrid deep learning pipelines," Neural Comput.Appl,vol.35,pp.22034–22048,2023.