Grenze International Journal of Engineering and Technology, June Issue

AI-Powered Browser Extension for Fake Job Post and Recruitment Scam Detection

B.L.S. Pranavi[1], D. Gnana Prasuna Reddy[2], K. Vasanth[3], A.K.S. Charan[4] and SK. Althaf Rahaman[5]
[1-4]Dept. of Computer Science and Systems Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, India
Email: pranavibommireddypalli@gmail.com, gnanaprasunadarapu@gmail.com, katarivasanth8@gmail.com, charanadidam1@gmail.com
[5]Assistant Professor, Dept. of Computer Science and Systems Engineering, Lendi Institute of Engineering and Technology, Vizianagaram, India
Email: rahmanalthaf@gmail.com

***Abstract*— Hire-Shield is an AI-powered system designed to detect and prevent Online Recruitment Fraud (ORF) by identifying fake job postings in real time. The system constructs and preprocesses a dataset of legitimate and fraudulent job postings, addressing class imbalance using the Synthetic Minority Oversampling Technique (SMOTE). Transformer-based NLP models such as BERT and RoBERTa are fine-tuned to capture subtle linguistic patterns characteristic of scam postings. These models are integrated into a Chrome browser extension that generates instant scam probability scores and highlights suspicious content directly on job portals. A supporting web dashboard provides analytics on fraud trends and multilingual accessibility. Hire- Shield enhances job seeker safety, reduces financial risk, and improves awareness of recruitment scams**

***Index Terms*— Online Recruitment Fraud, Fake Job Detection, Artificial Intelligence, NLP, Transformer Models, Chrome Extension, Scam Prevention.**

I. INTRODUCTION

Online recruitment has emerged as a dominant mechanism for talent acquisition across industries, offering unprecedented levels of accessibility for both employers and job seekers across the globe. Unfortunately, the rapid expansion of digital hiring platforms has also triggered an upward trajectory in a variant of cyber-enabled deception known as Online Recruitment Fraud. Fraudulent job postings are fitted with malignant scripts to fleece applicants by requiring upfront payments, personal information, or merely routing victims to phishing websites. These incidents affect students, fresh graduates, and unemployed individuals more, resulting in fiscal losses, identity theft, psychological trauma, and a lack of confidence in genuine recruitment ecosystems. Recent global cybersecurity and employment reports identify recruitment- related scams among the fastest-growing categories of online fraud, underlining an urgent need for scalable prevention mechanisms.

Traditional methods of addressing recruitment fraud depend to a great degree on job portals' manual moderation, employer verification processes, and user reports. Given the dynamic nature of the modus operandi by scammers, such methods, though effective up to a limit, are reactive, labor-intensive, and thus fail to keep pace with them.

Grenze ID: 01.GIJET.12.2.169_20

Most fraudsters manipulate language patterns, reuse legitimate company names, or create visually plausible postings that may get through platform-level filters. From a user's perspective, the typical red flags—such as unrealistic salaries or very unclear descriptions— are not always that apparent, especially for inexperienced applicants. This means that job seekers often come across fraudulent postings directly within trusted platforms without any indication of risk.

Several researchers have tried to address this challenge by investigating traditional machine learning techniques for the automated detection of fake jobs, including Support Vector Machines, Logistic Regression, and Random Forest classifiers trained on handcrafted features. Although these methods report moderate performance, they tend to rely highly on manual feature engineering and lack generalization across diverse job domains, platforms, and evolving scam patterns. Most rule-based and shallow learning methods cannot capture deep semantic cues hidden inside recruitment text, such as coercive language, urgency framing, or subtle inconsistencies in role descriptions.

Recent breakthroughs in artificial intelligence, specifically in NLP, have made transformer-based deep learning models reach the state of the art for text understanding tasks. Models like BERT and RoBERTa learn the contextual representations of language; hence, they can identify subtle linguistic patterns distinguishing fraud postings from nonfraud ones. On the contrary, the traditional classifiers rely on features engineered over the surface tokenization of job postings, whereas transformer models analyze an entire job description in a more holistic way, discerning implicit meanings and contextually relevant word relationships. At the same time, the deployment of such highly accurate models in real recruitment contexts faces challenges related to inference latency, user accessibility, and integrability into job browsing workflows.

In this work, we present Hire-Shield, an AI-driven browser- based system for detecting fake job postings and recruitment scams in real time. Our proposed approach combines a transformer-based NLP classification pipeline with lightweight Chrome extension that provides instant scam alerts within job portals. Job descriptions are extracted in real time, preprocessed to remove noise and normalize text, and passed through fine-tuned BERT and RoBERTa models that have been trained on a curated dataset of genuine and fraudulent postings. Due to the class imbalance commonly found in recruitment fraud data, the Synthetic Minority Oversampling Technique (SMOTE) is used in training the system. The system outputs a scam probability score, a safety verdict, and the highlights of suspicious phrases to enhance transparency and user trust.

The major contributions of this study are threefold:

- Design an end-to-end AI-driven framework for detecting online recruitment fraud using transformer-based NLP models
- Integrating the trained model into a browser extension that can perform real-time, user-centric scam detection during job browsing.
- Explainable output mechanism and analytics dashboard will be implemented for the purpose of increasing awareness, interpretability, and long- term monitoring of recruitment fraud trends. Hire- Shield combines deep semantic analysis with practical deployment to provide effective screening aids that enable job seekers to make better choices and reduce exposure to fraudulent recruitment.

II. LITERATURE SURVEY

Online recruitment fraud has emerged as a serious cybersecurity and socio-economic menace with the rapid proliferation of online recruitment platforms, job portals, and professional networking websites [1], [2]. Fraudulent job postings are increasingly designed to appear legitimate while covertly targeting applicants for monetary exploitation or sensitive personal information. Similar to other deceptive online threats, recruitment scams often employ subtle linguistic manipulation, urgency framing, and impersonation of trusted organizations [3]. With recruitment activities now largely shifted to digital ecosystems, there is a critical need for automated, accurate, and real-time fraud detection systems to protect vulnerable users, particularly students and early-career professionals [4], [5].

Early research in fake job posting detection primarily focused on rule-based systems and keyword-driven heuristics [6]. These approaches relied on manually defined red-flag indicators such as “registration fee,” “no interview required,” or “guaranteed placement,” which were hard-coded into detection frameworks. While such systems were effective against simple and repetitive scams, they lacked adaptability and failed to generalize to evolving fraud strategies. Their rigid design prevented semantic understanding of job descriptions, rendering them ineffective against sophisticated scam postings crafted to bypass rule-based filters [7].

To overcome these limitations, classical machine learning classifiers such as Support Vector Machines, Naïve Bayes, Logistic Regression, and Random Forests were introduced for automated recruitment fraud detection [8],

[9]. These models relied on handcrafted features extracted from job metadata, including company credibility indicators, salary anomalies, job description length, presence of external links, and communication channels. Several studies demonstrated that ensemble-based classifiers outperformed individual models [10]. However, these approaches remained heavily dependent on feature engineering and were unable to effectively capture the contextual semantics of recruitment text, resulting in limited robustness across platforms and job domains [11].

With advancements in Natural Language Processing (NLP), deep learning-based methods became increasingly dominant in text-based fraud detection. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) architectures, were used to model sequential dependencies within job descriptions, while Convolutional Neural Networks (CNNs) were applied to extract localized textual patterns associated with scam intent [12], [13]. Zhang et al. demonstrated that sequence-aware neural models outperform traditional machine learning approaches by learning stylistic and contextual scam patterns [14]. However, these architectures struggled with long-range semantic dependencies and multilingual recruitment content, limiting their effectiveness in real-world applications [15].

The emergence of transformer-based models such as BERT and RoBERTa marked a significant shift in recruitment fraud detection research [16], [17]. These models employ self-attention mechanisms to analyze entire job descriptions holistically, enabling deeper understanding of implicit fraud indicators and contextual cues. Multiple studies have reported substantial improvements in classification accuracy when transformers were fine-tuned on labeled recruitment datasets [18]. Al-Ameen etal. demonstrated accuracy exceeding 95% using BERT for employment scam detection, highlighting the superiority of transformer-based semantic modelling over earlier techniques [19].

Beyond model architecture, dataset quality and class imbalance remain critical challenges in recruitment fraud detection. Legitimate job postings significantly outnumber fraudulent ones, resulting in biased models with poor scam sensitivity. To address this issue, resampling techniques such as the Synthetic Minority Oversampling Technique (SMOTE) have been widely adopted to improve minority class representation [20].

Additionally, preprocessing steps including tokenization, text normalization, stop-word removal,and noise filtering have been shown to enhance modelstability and generalization, particularly for transformer-based training [21].

Despite improved detection accuracy, much of the existing literature remains limited to offline experimentation and lacks real-world deployment considerations. The absence of real-time inference, weak user feedback mechanisms, and limited explainability restrict the practical usability of these systems [22]. Recent research has therefore shifted toward browser-integrated approaches that provide instant alerts, interpretable outputs, and user-centric interfaces [23], [24]. Such systems aim to seamlessly integrate with job portals, enabling users to receive scam warnings during active browsing rather than post-application analysis.

Collectively, existing studies demonstrate a clear transition from rule-based filtering and classical machine learning models to deep learning-driven semantic analysis using transformer architectures. While transformers significantly enhance scam detection accuracy and contextual understanding, challenges remain in handling data imbalance, adapting to evolving fraud strategies, and presenting interpretable outputs to end users. Furthermore, most existing systems function as back-end services and lack direct integration into the job-seeking workflow. The proposed Hire-Shield system addresses these gaps by integrating SMOTE-based dataset engineering with fine- tuned BERT and RoBERTa models and deploying them through a Chrome extension for real-time scam detection. The inclusion of explainable outputs and an analytics dashboard further improves detection effectiveness, transparency, and user trust, making Hire- Shield a practical and scalable solution for online recruitment fraud prevention.

III. METHODOLOGY

The proposed framework integrates transformer-based text classification with real-time deployment through a browser extension. The methodology consists of dataset preparation, model training, evaluation, and deployment.

A. Dataset Description

The dataset used in this study was constructed by merging two publicly available datasets of recruitment job postings. The first dataset contained 17,880 labelled job postings (legitimate and fraudulent), and an additional 10,000 fraudulent job postings were incorporated to enhance scam pattern diversity. After aligning common attributes, the merged dataset contained 27,880 job postings across 10 structured features.

The selected attributes included: title, description, requirements, company_profile, location, salary_range, employment_type, industry, benefits, and a binary fraudulent label.
The final class distribution consisted of 17,014 legitimate (0) samples and 10,866 fraudulent (1) samples.

B. Text Preprocessing

Raw job posting data were cleaned and standardized prior to model training. HTML tags, URLs, special characters, and redundant whitespace were removed, and all text was converted to lowercase. Several attributes contained missing values, particularly salary_range (53.85%), benefits (25.87%), and industry (17.59%). These were handled by replacing null values with empty strings to maintain input consistency. The cleaned textual content was then tokenized using the BERT/RoBERTa tokenizer, which converts text into subword tokens and corresponding input IDs suitable for transformer-based classification.

C. Class Imbalance Handling using SMOTE

The merged dataset contained 27,880 records with 17,014 legitimate and 10,866 fraudulent postings, indicating moderate class imbalance. To improve minority-class sensitivity during training, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set. SMOTE generates synthetic fraud samples by interpolating between nearest neighbors. SMOTE generation equation

$$X\_\{new\} = x\_i + \lambda(x\_\{nn\} - x\_i), \lambda \sim U(0,1), \quad (1)$$

where x _i represents a minority class sample, x_nn denotes one of its nearest neighbors, and λ is a random value drawn from a uniform distribution between 0 and 1.

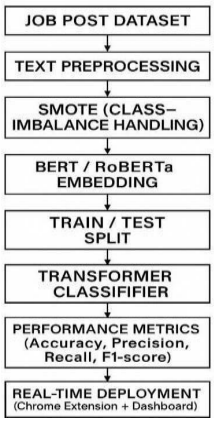


Fig.1. Proposed Architecture

D. Transformer-based Text Embedding

The cleaned job postings were encoded using the pre-trained bert-base-uncased model. The textual fields (title, description, requirements, and company_profile) were concatenated into a single input sequence and tokenized using the BERT tokenizer. Contextual embeddings were generated through transformer encoder layers, and the final hidden representation corresponding to the [CLS] token was passed to a linear classification layer for binary prediction. The [CLS] representation was passed to a linear layer to predict the probability of a job posting being fraudulent. The model was optimized using AdamW with Binary Cross-Entropy loss.

E. Train–Test Split

The dataset was divided using an 80:20 stratified split, resulting in 22,304 training samples and 5,576 testing samples. The split ensured evaluation on previously unseen job postings to assess generalization performance.

F. Training Configuration

Model fine-tuning was performed for 2 epochs using GPU acceleration (CUDA) in Google Colab. The final training loss converged to 0.0713, and validation accuracy reached 98.8%.

G. Performance Evaluation

Model performance was evaluated using Accuracy, Precision, Recall, and F1-score on the test dataset to assess fraud detection effectiveness.

H. Real-Time Deployment

The trained BERT-based classifier was integrated into a Chrome browser extension for real-time scam detection. When a user visits a job listing, the extension extracts relevant textual fields and sends them to the backend model for inference. The system returns a probability score along with a risk classification (High Risk, Suspicious, or Likely Real), which is displayed directly within the browser interface. The analytics dashboard aggregates scan statistics, detection rates, and recent classification logs to provide transparency and monitoring of fraud trends.

IV. Results and Discussion

The proposed transformer-based fake job detection system was evaluated on a held-out test set comprising 5,576 job postings using an 80:20 stratified split. Model performance was assessed using Accuracy, Precision, Recall, F1-score, Receiver Operating Characteristic (ROC) curve.

A. Quantitative Results

The fine-tuned BERT (bert-base-uncased) classifier achieved the following performance on the test dataset.

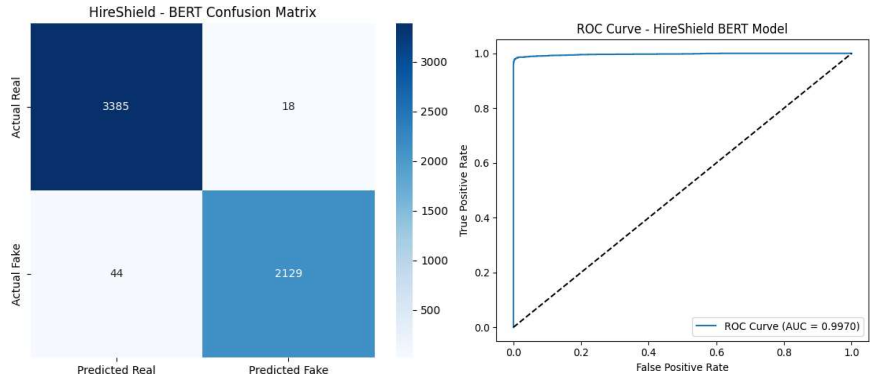
Table I: Performance of the Proposed Bert-based Classifier

Metric	Value
Accuracy	98.89%
Precision (Fake)	99.16%
Recall (Fake)	97.97%
F1-Score	98.56%
AUC (ROC)	0.9970
Average Precision (PR)	0.9968

The proposed model achieved an accuracy of 98.89%, with fraud precision of 99.16% and fraud recall of 97.97%. The F1-score for the fraudulent class was 98.56%. These results indicate strong classification capability with balanced precision and recall.

The confusion matrix is presented in Fig. 2.

The model correctly classified 5,514 out of 5,576 samples. False positives (18 cases) represent legitimate postings incorrectly flagged as fraudulent, while false negatives (44 cases) correspond to missed fraudulent postings. The relatively low false negative count demonstrates strong fraud detection sensitivity.



(a) Confusion matrix (b)ROC Curve

Fig.2. Performance evaluation of the proposed BERT-based fake job detection model: (a) Confusion matrix showing classification outcomes on the test set; (b) Receiver Operating Characteristic (ROC) curve with AUC = 0.997.

The classifier achieved an Area Under the Curve (AUC) of 0.9970, indicating near-perfect class separability across varying decision thresholds.

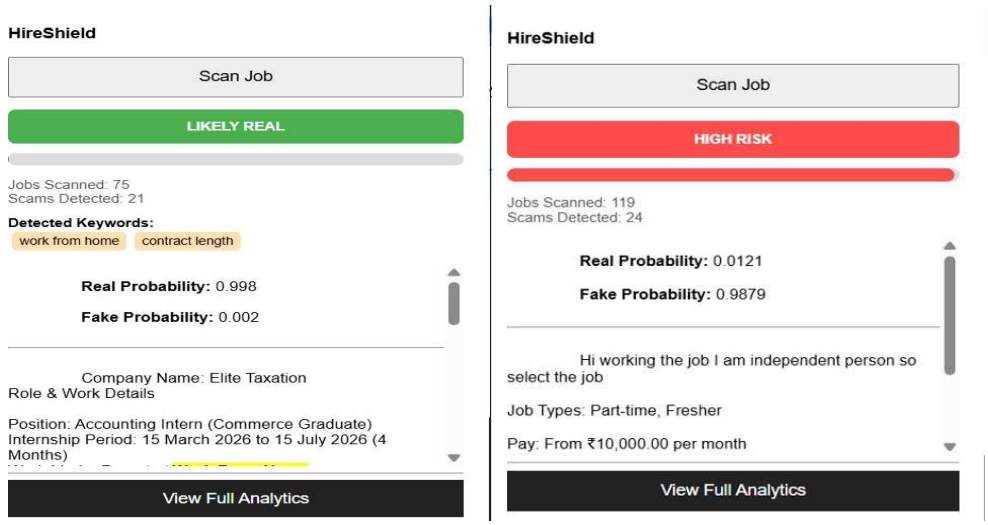
B. Baseline Comparison

To evaluate the benefit of contextual embeddings, a TF-IDF + Logistic Regression baseline model was implemented. The baseline achieved 98.26% accuracy, with 100% precision for the fraudulent class but 95.53% recall. The confusion matrix revealed 97 false negatives and zero false positives. In comparison, the BERT-

based classifier reduced false negatives from 97 to 44 and improved fraud recall to 97.97%, detecting 53 additional fraudulent postings. Although 18 false positives were introduced, the reduction in undetected scams demonstrates improved sensitivity to nuanced fraud patterns.

C. Qualitative Results

The trained model was deployed through a Chrome browser extension for real-time scam detection. Representative outputs are shown in Fig. 4. Scam-style postings with urgency manipulation and informal recruiter contact details were assigned fraud probabilities exceeding 0.999, while legitimate corporate job postings received real-job probabilities above 0.99.



(a) Likely Real (b) High Risk

Fig.3. Real-time output of the HireShield browser extension: (a) legitimate job ; (b) fraudulent job

D. Discussion

The results demonstrate that both lexical and contextual features are predictive of recruitment fraud. While the logistic regression baseline performed strongly due to explicit fraud-related keywords, the transformer-based classifier exhibited superior fraud recall and threshold robustness. The BERT model reduced missed scams by approximately 55% compared to the baseline. This highlights the effectiveness of contextual embeddings in detecting subtle or creatively written fraudulent postings. Despite the strong performance, highly sophisticated scam postings that closely resemble legitimate corporate advertisements may occasionally evade detection. Future improvements may include integration of behavioral metadata, recruiter domain verification, multilingual extension, and explainable AI mechanisms.

V. CONCLUSION

This study presented an AI-powered browser extension for real-time detection of fraudulent job postings using transformer based natural language processing. The system integrates a fine-tuned BERT (bert-base-uncased) classifier with SMOTE- based class balancing and Chrome-based deployment to analyze job descriptions and provide immediate scam risk assessments. The proposed model achieved 98.89% accuracy with an AUC of 0.997 on a held-out test set of 5,576 job postings. Confusion matrix analysis demonstrated strong fraud detection capability, reducing false negatives to 44 cases while maintaining minimal false positives. Compared to a TF-IDF + Logistic Regression baseline, the transformer-based approach improved fraud recall and detected 53 additional fraudulent postings, highlighting the benefits of contextual semantic modeling. Deployment within a browser extension demonstrates the practical applicability of the framework. Real- time testing confirmed consistent alignment between predicted probabilities and common scam indicators, including urgency framing, unrealistic compensation claims, and suspicious contact details. Despite strong performance, certain limitations remain. The current approach relies primarily on textual features and does not incorporate recruiter behavioral metadata or domain verification signals. Future work may explore explainable AI mechanisms, multilingual

expansion, adaptive learning strategies, and broader institutional integration. Overall, the findings demonstrate that transformer-based contextual language models provide a scalable and effective solution for mitigating online recruitment fraud and improving digital job-search safety.

REFERENCES

[1] K. Taneja, J. Vashishtha, and S. Ratnoo, "Fraud-BERT: Transformer-based context-aware online recruitment fraud detection," Discover Computing, vol. 28, no. 9, pp. 1–26, 2025.

[2] P. A. Kumar, G. Likitha, B. Nithin, A. Akash, and G. Saikumar, "Detection of fake online recruitment using machine learning," International Journal of Scientific Development and Research, vol. 9, no. 5, pp. 112–118, May 2024.

[3] S. Banerjee, R. Dutta, and M. Kundu, "Fake job post detection using machine learning approaches," International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 45–50, 2020.

[4] M. Abbas, R. Sameh, and Y. A. Ali, "Online recruitment fraud detection using deep learning," International Journal of Research Publication and Reviews, vol. 6, no. 6, pp. 211–218, 2025.

[5] R. Dada, S. Bassi, and A. Patel, "Online recruitment fraud detection system," International Journal of Engineering Scientific Research, vol. 12, no. 2, pp. 34–40, 2021.

[6] V. Sharma and S. Gupta, "Fake job posting detection using NLP and machine learning," International Journal of Advanced Studies in Engineering and Technology, vol. 8, no. 1, pp. 89–95, 2022.

[7] R. Singh and K. Verma, "Detecting real versus fake job postings using artificial neural networks," International Journal of Scientific Development and Research, vol. 10, no. 3, pp. 301–307, 2025.

[8] A. Akram, R. Irfan, and A. S. Al-Shamayleh, "Online recruitment fraud detection using deep learning approaches," IEEE Access, vol. 12, pp. 45678–45689, 2024.

[9] S. Pillai, "Detecting fake job postings using bidirectional LSTM," arXiv preprint arXiv:2304.02019, Apr. 2023.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[11] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[12] 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.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[14] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of fraud patterns in text data," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[15] Google, "Google Chrome adds AI-powered scam detection to keep users safe," TechGig, Dec. 2024. [Online]. Available: https://content.techgig.com/technology/google-chrome-adds-ai-powered-scam-detection-to-keep-you-safe/articleshow/116623300.cms. Accessed: Mar. 2025.

[16] Google, "Common Resolve – AI-based scam detection extension," Chrome Web Store. [Online]. Available: https://chromewebstore.google.com/detail/common-resolve/hemkacfkmhldlhpfndjfibkalfphiibf. Accessed: Mar. 2025.

[17] H. Zhang, Z. Wang, and Y. Liu, "Text-based fraud detection using attention mechanisms," Computers & Security, vol. 110, pp. 102–118, 2022.

[18] Omdena, "Detecting fake job postings using NLP," Omdena Project Report, 2023. [Online]. Available: https://www.omdena.com/projects/detecting-fake-job-postings-using-nlp.

[19] S. Ramesh and A. Karthik, "Fake job post detection using ensemble machine learning models," International Journal of Modernization in Engineering, Technology and Science, vol. 5, no. 2, pp. 98–105, 20