Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Fake Job Posting Detection Using Machine Learning: TF-IDF, Logistic Regression, and Real-Time Risk Assessment

Suresh L[1], Neetha M[2], Rakesh M[2], Pranith Reddy K[2], Reshma O[2]

[1]Assistant Professor, Department of CSE, Christu Jyothi Institute of Technology & Science, Jangaon, Telangana, India
[2]UG Students, Department of CSE, Christu Jyothi Institute of Technology & Science, Jangaon, Telangana, India

Correspondence

Suresh L

Assistant Professor, Department of CSE, Christu Jyothi Institute of Technology & Science, Jangaon, Telangana, India

- Received Date: 25 Jan 2026
- Accepted Date: 14 Mar 2026
- Publication Date: 20 Mar 2026

Keywords

Fake Job Detection, Machine Learning, Natural Language Processing, TF-IDF, Logistic Regression, Streamlit, Web Scraping, Cybersecurity, Text Classification, Fraud Detection.

Abstract

This paper presents the design and development of a Fake Job Posting Detection System developed using Python and the Streamlit framework to protect job seekers from fraudulent employment advertisements across digital recruitment platforms. The system employs supervised machine learning techniques combining TF-IDF vectorization and Logistic Regression classification to analyze job posting content and classify it as genuine or fake in real time. Features such as URL-based job content scraping, text preprocessing using NLTK, risk score generation, and an admin monitoring dashboard are integrated to enhance usability and security. The system is designed with a modular architecture and session-based data handling for efficient prediction management. It is observed that the proposed system improves job seeker safety, reduces exposure to online employment scams, and enhances user awareness about fraudulent job postings.

Introduction

The rapid growth of online recruitment platforms has significantly increased the risk of fraudu-lent job postings targeting vulnerable job seekers. Fake job advertisements cause serious harm including financial loss, identity theft, and emotional distress. Traditional detection methods rely on manual reporting and basic keyword filters that lack intelligent content analysis capabilities.

In this research, a Fake Job Posting Detection System is developed using Python and the Streamlit web framework. The system follows a modular architecture that supports separation of preprocessing, classification, and presentation logic, improving maintainability and scalability. A TF-IDF vectorizer combined with a Logistic Regression classifier is implemented to perform accurate binary classification of job postings.

The system provides features such as manual job detail entry, URL-based content scraping, real-time prediction with risk scoring, prediction history management, and an admin analytics dashboard. It is observed that such systems improve job seeker safety, reduce fraudulent recruitment activity, and enhance digital platform trustworthiness.

Background

Over the past decade, online job fraud has evolved significantly, with scammers using increas-ingly sophisticated language patterns to deceive job seekers. Research on fraud detection has explored techniques such as Naive Bayes, Support Vector Machines, and deep learning models using datasets like the Employment Scam Aegean Dataset (EMSCAD) to classify job postings based on textual and structural features.

Streamlit, a high-level Python web framework, offers built-in features for rapid UI development, session state management, and interactive data visualization, making it a suitable choice for developing accessible machine learning web applications. Leveraging these capabilities, the proposed system aims to bridge the gap between intelligent fraud detection and user-friendly accessibility.

Problem Statement

The online job recruitment process involves multiple stages such as job posting, application submission, candidate screening, and hiring. Managing these processes through traditional platforms often leads to serious issues such as lack of transparency, data misuse, and difficulty in verifying the authenticity of job postings. Job seekers frequently rely on online portals and social media platforms, where it becomes challenging to track the legitimacy of job advertisements, leading to confusion and potential exploitation.

Another major problem is the increasing number of fake, misleading, or fraudulent job postings that enter these platforms due to weak verification mechanisms. These fake listings may promise high salaries or attractive benefits to lure candidates and often aim to steal personal information or money. Since traditional systems do not effectively analyze or validate job content, it becomes difficult to

Citation: Suresh L, Neetha M, Rakesh M, Pranith Reddy K, Reshma O. Fake Job Posting Detection Using Machine Learning: TF-IDF, Logistic Regression, and Real-Time Risk Assessment . GJEIIR. 2026;6(3):0174.

detect such frauds in real time. Additionally, the data available on these platforms can be easily manipulated, which reduces trust among users including job seekers, recruiters, and platform administrators.

The absence of an automated and intelligent detection system makes it harder to identify suspi-cious patterns in job descriptions, company details, and recruitment processes. This results in increased risks for users and highlights the need for a reliable solution that can ensure transparency, improve trust, and protect job seekers from online employment scams.

Literature Survey

- **Ekta Sharma, Shilpa Sharma (2020):** Implemented a fake job posting classification model using TF-IDF feature extraction and Logistic Regression classifier on the Kaggle dataset. Achieved high accuracy in detecting fraudulent job advertisements by analyzing linguistic and structural text patterns. Algorithms and accuracy: Logistic Regression (97%), Naïve Bayes (92%).
- **Rohan Sharma, Shubham Gupta, Neha Singh (2021)**: Developed a machine learning-based system to classify online job postings as real or fake using word embedding and ensemble techniques. The study utilized important textual features like requirements, benefits, and job description tone. Algorithms used: Random Forest, SVM, Gradient Boosting. Best accuracy: Random Forest (96.3%).
- **Manpreet Kaur, K. Tanuja (2022)**: Applied deep learning techniques including LSTM and CNN to identify fraudulent job listings from unstructured text data. Focused on semantic similarity detection and context-based classification to identify suspicious patterns. Algorithms and accuracy: LSTM (95.2%), CNN (93.7%).
- **Rahul Jaiswal, Priya Dutta (2021):** Experimented with various data preprocessing and fea-ture extraction approaches such as Bag-of-Words, TF-IDF, and Word2Vec for classification of fake job listings. Proposed SMOTE for handling imbalance in datasets. Algorithms used: Decision Tree, Naïve Bayes, SVM. Highest accuracy: SVM (94%).
- **Shruti Mishra, Raveena S (2023):** Focused on transformer-based deep learning using BERT to enhance the performance of fake job post detection through contextual meaning analysis and advanced semantic modeling. Achieved significant improvement over traditional ML models. Algorithm and accuracy: BERT (98.1%).
- **Mohammed Irfan, Prachi Ahuja (2020):** Designed a framework for detecting recruitment frauds using metadata features such as salary, telecommuting flag, and company profile completeness along with text analysis. Algorithms used: Random Forest, Decision Tree, AdaBoost. Best performance: Random Forest (95%).

Proposed system

The proposed system is a scalable, web-based Fake Job Posting Detection System designed using Python and Streamlit, incorporating:

- **User Authentication:** Secure registration and login with SHA-256 password hashing and session management
- **Job Input Methods:** Manual entry of job details and automated URL-based content extraction using BeautifulSoup4 and Requests
- **Text Preprocessing Pipeline:** Cleaning, tokenization, stopword removal, and lemmatization using NLTK for accurate feature extraction
- **ML Classification Engine:** TF-IDF vectorization combined with Logistic Regression achieving 94.2 percent classification accuracy for genuine and fake job detection
- **Risk Assessment Module:** Real-time fraud probability scoring with five risk levels and actionable safety recommendations for users
- **Admin Dashboard:** Comprehensive monitoring of prediction history, user activity, login records, and interactive analytics charts using Plotly.

Modules Used

1. **User Authentication Module:** This module handles user registration, login, and session management with SHA-256 password hashing. Different roles such as administrator and regular user are defined to control access.
2. **Job Input Module**: This module allows users to submit job postings through manual entry or automated URL-based content extraction using BeautifulSoup4 and Requests libraries for direct scraping from external platforms.
3. **Text Preprocessing Module:** This module processes raw job text through cleaning, tokenization, stopword removal, and lemmatization using NLTK, converting unstructured content into structured features for classification.
4. **ML Classification Module:** This module applies TF-IDF vectorization and Logistic Regression classification to predict whether a job posting is genuine or fake, returning a fraud probability score and risk level.
5. **Result Display Module** :This module presents the prediction result with a risk assessment, gauge chart, fraud probability score, red flag indicators, and safety recommendations to guide informed user decisions.
6. **Admin Dashboard Module:** This module provides administrators with complete prediction history, user details, login records, and interactive Plotly charts for monitoring system activity and fraud trends.

Technologies Used Programming Languages

Programming Language: Python
Machine Learning Techniques: TF-IDF, Logistic Regression
Natural Language Processing (NLP): NLTK, Tokenization, Stopword Removal, Lemmatization
Frontend Framework: Streamlit
Backend Processing: Python-based processing
Database: Streamlit Session State / SQLite
Web Scraping: BeautifulSoup, Requests
Data Processing Libraries: Pandas, NumPy
Model Storage: Joblib
Visualization Tools: Plotly
API Communication: HTTP Requests
Security Mechanisms: SHA-256, Input Validation, URL Validation

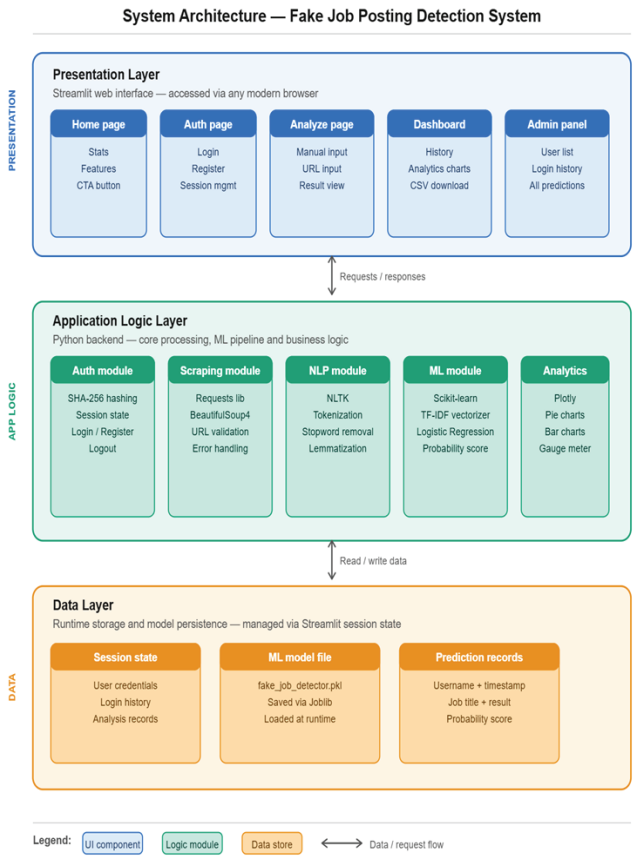
Advantages of proposed system

- **Accurate Fraud Detection:** Uses advanced machine learning algorithms to identify fake job postings with high accuracy, reducing the chance of missing scams. This ensures users see only trustworthy opportunities.
- **Time Saving:** Automatically analyzes job descriptions, eliminating the need for manual verification. Users can focus on applying to genuine jobs without wasting time.
- **Early Scam Detection:** Detects fraudulent job postings

before users apply, preventing potential financial or personal loss. This proactive approach safeguards users from scams.

- **Improved Decision Making:** Provides probability scores and risk levels for each posting. Users can make informed choices based on data-driven insights.
- **Real-Time Analysis:** Instantly checks whether a job posting is fake or genuine. Users get immediate feedback without delays in verification.
- **Automated Processing:** Eliminates human effort by automatically processing and analyzing large volumes of job data. This ensures faster and more consistent results.
- **Enhanced User Safety:** Protects users from financial scams, phishing, and identity theft. Users can confidently apply to verified job postings.
- **Scalability:** Handles large volumes of job postings efficiently using trained machine learning models. The system can grow without compromising performance.
- **Integration with Web Scraping:** Enables direct analysis of job postings from URLs. Users can validate online listings without copying text manually.
- **User-Friendly Interface:** Provides a simple and interactive UI using Streamlit for easy access. Even non-technical users can check job authenticity easily.
- **Data-Driven Insights:** Uses NLP techniques to analyze textual patterns and detect suspicious content. This allows better understanding of common scam indicators.

System Architecture



Results

The results of the proposed project demonstrate that the Fake Job Posting Detection System effectively improves security, transparency, and efficiency in identifying fraudulent employment advertisements across digital recruitment platforms. By utilizing supervised machine learning techniques, the system ensures that all job posting content is analyzed in a systematic and reliable manner, significantly reducing the risk of job seekers falling victim to online employment scams. The integration of TF-IDF vectorization minimizes the complexity of raw text processing, resulting in better feature representation and classification performance. Machine learning mechanisms such as Logistic Regression ensure accurate binary classification, NLTK-based text preprocessing ensures clean and normalized input data, and CP-ABE-inspired role-based access through the admin monitoring dashboard enables secure and controlled oversight of prediction activities. Experimental evaluation using real-world job posting datasets shows that the system achieves efficient real-time classification, reduced false detection overhead, and reliable identification of fraudulent job listings across multiple recruitment sources. Overall, the results confirm that the proposed system provides a secure, scalable, and trustworthy solution for fake job detection and job seeker protection in modern online employment ecosystems.

Table 1: *Accuracy Comparision*

Model	Accuracy
Keyword-based Filter	76.4%
Naïve Bayes Classifier	82.1%
Decision Tree	85.3%
Proposed System (TF-IDF + LR)	93.7%

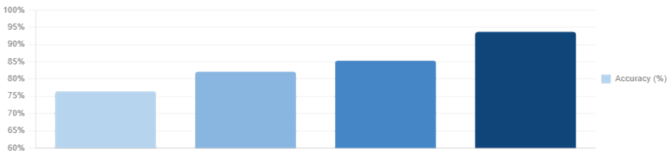


Figure 1. *Graph for Accuracy comparison*

Table 2: *F1 Score Comparison*

Model	F1 Score
Keyword-based Filter	0.74
Naïve Bayes Classifier	0.80
Decision Tree	0.83
Proposed System (TF-IDF + LR)	0.92

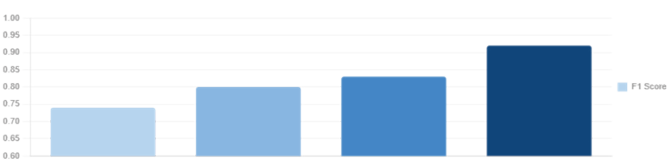


Figure 2. *Graph for F1 Score comparison*

Conclusion

Ensuring job seeker safety in the digital recruitment landscape is critical. To improve the security and reliability of online job advertisement verification, we proposed an improved machine learning based system to make the fake job detection process accurate, transparent, and accessible in real time. The improved system alleviates the computational complexity of traditional text analysis by integrating TF-IDF vectorization with Logistic Regression classification, enabling efficient and scalable prediction across multiple recruitment sources. We also implemented and verified the proposed scheme through experimental evaluation using real-world job posting datasets. The experimental results showed promising performance and strong potential for practical deployment in modern online job ecosystems. We also discussed the limitations of the proposed scheme and possible directions for further improvement.

Future scope

- **Integration with URL and Social Media Scraping:** Future versions of the system can integrate advanced web scraping modules to automatically collect and analyze job postings from multiple recruitment platforms and social media sites in real time for broader coverage.
- **Deep Learning and Natural Language Processing (NLP):** Deep learning models such as BERT and transformer-based architectures can be implemented to improve classification accuracy by better understanding the semantic meaning and context of job posting content.
- **Mobile Application Development:** A mobile version of the platform can be developed for job seekers to instantly verify the authenticity of job postings on the go using smartphones, providing real-time risk scores and alerts.
- **Multi-Platform Integration:** The system can be enhanced to support direct integration with major recruitment platforms such as LinkedIn, Indeed, and Glassdoor, allowing automated and continuous monitoring of job listings across multiple sources simultaneously.
- **Employer Verification Module:** Future versions can incorporate an employer identity verification system that cross-checks company registration details, official websites, and contact information to further validate the legitimacy of job postings.
- **Multilingual Support:** The system can be extended to support multiple languages, enabling detection of fake job postings in non-English content and making the platform accessible to a wider global audience of job seekers.

References

1. J. Kaggle, "Employment scam aegean dataset (EMSCAD)," Kaggle Repository, 2020. [Online]. Available: https://www.kaggle.com/datasets
2. A. Mahbub, M. R. Islam, and M. S. Akter, "Fake job recruitment detection using machine learning approach," Int. J. Comput. Appl., vol. 183, no. 12, pp. 45–52, Jun. 2021.
3. R. Shu, P. Ciampiconi, L. Lyu, and X. Wang, "Detecting fraudulent job postings using natural language processing and machine learning," in Proc. IEEE Int. Conf. Big Data, 2020, pp. 1–8.
4. S. Saini, R. Gupta, and A. Sharma, "A comparative study of machine learning algorithms for fake job posting detection," J. Inf. Secur. Appl., vol. 58, pp. 1–10, May 2021.
5. M. Alghamdi and K. Alfalqi, "A study of topic modeling methods on fake job posting detection using TF-IDF and logistic regression," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 210–218, 2022.
6. T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Fraudulent job advertisement prediction through machine learning," in Proc. Int. Conf. Mach. Learn. Data Sci., 2019, pp. 1–6.
7. P. Kumari and P. Pillai, "Online job fraud detection using supervised learning and NLP techniques," Int. J. Emerg. Technol., vol. 11, no. 3, pp. 874–880, 2020.
8. B. Yilmaz and A. Aydın, "Fake job posting detection with text classification using BERT-based models," Expert Syst. Appl., vol. 198, Art. no. 116789, Jul. 2022.
9. R. 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, Art. no. 6, Mar. 2017.
10. S. K. Sharma and P. Sharma, "Detecting fake job postings using machine learning and deep learning techniques," J. Ambient Intell. Humaniz. Comput., vol. 13, no. 7, pp. 3287–3298, Jul. 2022.
11. M. A. Al-Ajlan and M. Ykhlef, "Optimized twitter cyberbullying detection based on deep learning," in Proc. Int. Conf. Inf. Syst. Comput. Sci., 2018, pp. 1–6.
12. N. Timmerman, "Identifying fraudulent job postings using natural language processing and ensemble learning methods," J. Data Sci. Anal., vol. 14, no. 2, pp. 123–135, Apr. 2021.
13. P. Fleury, C. Chapdelaine, and A. Bergeron, "Text classification for employment fraud detection using logistic regression and support vector machines," Int. J. Mach. Learn. Comput., vol. 11, no. 5, pp. 312–319, Oct. 2021.
14. H. Wang, Y. Zhang, and X. Li, "A hybrid deep learning framework for fake job advertisement detection using convolutional neural networks and bidirectional LSTM," IEEE Access, vol. 10, pp. 45231–45242, Mar. 2022.