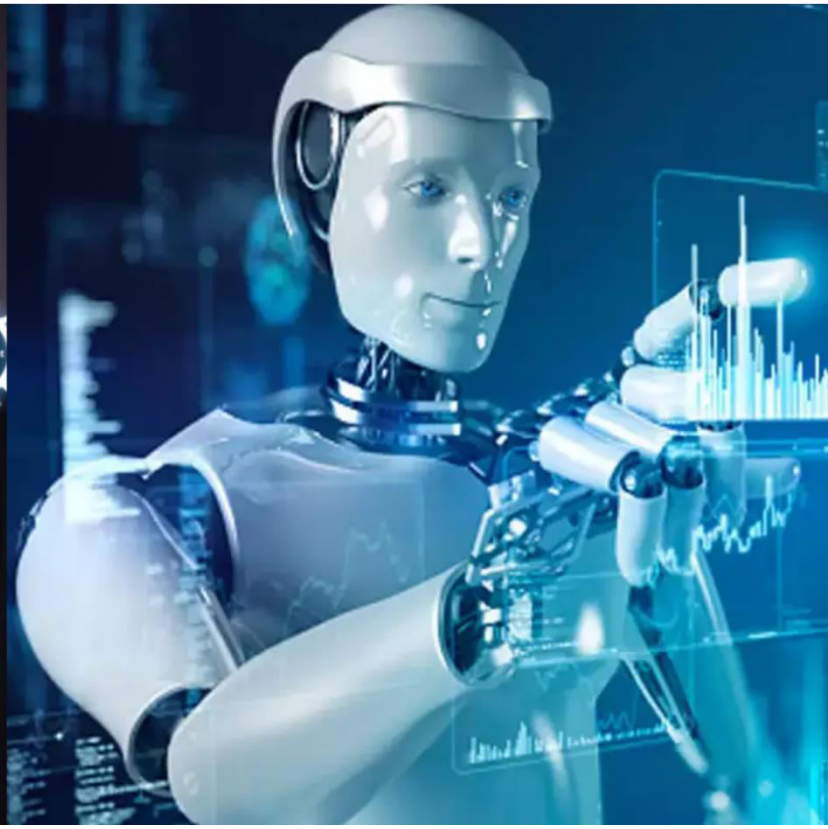
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 6, June 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406078|

Fake Job Post Detection using Machine Learning

Kadamanchi Revathi[1], Kalakota Saniya[2], G. Naga Sujini[3], Dr. K. Rajitha [4], R. Mohan Krishna [5]

Student, Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India[1,2]

Assistant Professor, Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India[3,,4,5]

ABSTRACT: The rise of online job platforms has improved the hiring process but also led to an increase in fake job advertisements, posing risks to job seekers. This project presents a Machine Learning-based system for automatic fake job post detection. A labeled dataset of real and fake job descriptions was processed using Natural Language Processing (NLP) techniques. TF-IDF vectorization converted the text data into numerical form, and a Logistic Regression model was trained for binary classification. A web application was developed using Flask to provide real-time prediction. Users can input job URLs, and the system scrapes the content using Requests and BeautifulSoup4, then classifies it as real or fake. The approach is accurate, scalable, and user-friendly, demonstrating the effective integration of ML and NLP to enhance online recruitment security. Future enhancements include dataset expansion and deep learning model integration.

KEYWORDS:-Fake Job Detection Machine Learning, NLP, Logistic Regression, TF-IDF, Web Scraping, Flask, Job Post Analysis, BeautifulSoup, Requests, Scikit-learn, Feature Extraction, Content Moderation, Python.

I. INTRODUCTION

Online job portals have revolutionized the hiring process, providing seamless interaction between employers and job seekers. However, their growing popularity has also made them targets for cybercriminals who post fake job advertisements to deceive users, steal sensitive information, or execute financial scams. Manual detection of such listings is time-consuming and unreliable, necessitating automated solutions. This study presents a system that leverages Machine Learning (ML) and Natural Language Processing (NLP) to detect fraudulent job postings. A labeled dataset comprising genuine and fake job descriptions was used, with TF-IDF vectorization applied for feature extraction. A Logistic Regression (LR) model was trained for binary classification. Additionally, a user-friendly web application was developed using Flask, integrating web scraping tools (Requests and BeautifulSoup4) to analyze content from user-submitted URLs. The proposed system is fast, scalable, and effective, offering a practical solution to a critical cybersecurity issue in the digital employment space.

II. LITERATURE SURVEY

Sharma, P. & Verma, K. in their study [1] proposed an ensemble model integrating Decision Trees, Random Forests, and Gradient Boosting for fake job post detection. Their method demonstrated enhanced accuracy and robustness compared to individual classifiers. However, the increased computational load and training time introduced scalability concerns.

Ahmed, L. & Singh, A. in [2] explored the application of the BERT model with fine-tuning techniques to detect fraudulent job postings. Their model achieved an impressive F1-score of 93%, showcasing the effectiveness of transformer-based NLP models. However, the need for domain expertise and high computational power to avoid overfitting was a noted limitation.

Johnson, M. & Liu, Y. in [3] emphasized the role of feature engineering for fake job detection. By extracting attributes such as suspicious keywords, email domains, and job title patterns, they trained a Logistic Regression model with strong performance. Yet, the reliance on manual feature design reduced scalability and adaptability.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406078|

Chen, H. & Wu, X. in [4] used traditional text mining techniques including TF-IDF and n-gram features with a Support Vector Machine (SVM) classifier, achieving 91% accuracy. The approach was efficient but lacked the deeper contextual understanding provided by modern NLP architectures like BERT.

Wang, X. et al. in [5] implemented the BERT model for detecting fake job advertisements. Leveraging BERT's contextual depth, their model achieved an F1-score of 0.92. While highly accurate, the method demanded a large labeled dataset and significant computational resources, making deployment in low-resource settings challenging.

Thomas, R. & Kumar, A. in [6] proposed an ensemble model combining Random Forest and XGBoost to enhance detection accuracy. This method demonstrated improved robustness but required careful hyperparameter tuning and was computationally intensive, limiting its use in real-time applications.

Singh, A. & Gupta, M. in [7] applied a Bi-LSTM (Bidirectional Long Short-Term Memory) model for employment scam detection. The model's strength lay in its ability to capture sequential text patterns, but it required extensive labeled data and long training times, affecting scalability.

Lee, K. et al. in [8] utilized an LSTM-based model that achieved 97% accuracy in identifying fraudulent job postings. The model's ability to learn complex text patterns contributed to its success. However, the high computational cost and slower training process posed deployment challenges in practical scenarios.

III. PROPOSED METHODOLOGY AND DISCUSSION

3.1 System Architecture:

The proposed system for fake job post detection is designed as an end-to-end, modular pipeline integrating data collection, natural language processing (NLP), machine learning (ML), and real-time feedback mechanisms. As shown in Fig. 1, the architecture begins with the data collection and ingestion module, which
gathers raw job posting data from various sources including web scrapers, public APIs, and social media platforms.

The collected raw data is then passed through a data preprocessing pipeline that includes text tokenization, Named Entity Recognition (NER), and feature extraction. These processed features are used to train and deploy a supervised machine learning model, incorporating algorithms such as Logistic Regression, XGBoost, and BERT to classify job advertisements as fake or legitimate.

Predictions generated by the model are evaluated in the real-time job post scoring module, which flags potentially fake posts. Flagged content is routed through an automated removal workflow, which either removes the posts through API calls or escalates them for manual review when necessary.

A critical component of the architecture is the feedback loop and continuous learning module. This unit incorporates user feedback and newly flagged posts to periodically retrain the model, thereby improving prediction accuracy over time. Additionally, the system includes monitoring and reporting functionalities, which log audit trails, flag alerts, and capture key performance indicators via a metrics dashboard.

This architecture ensures scalability, adaptability, and real-time responsiveness, making it a robust solution for identifying and managing fake job advertisements across digital employment platforms.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406078|

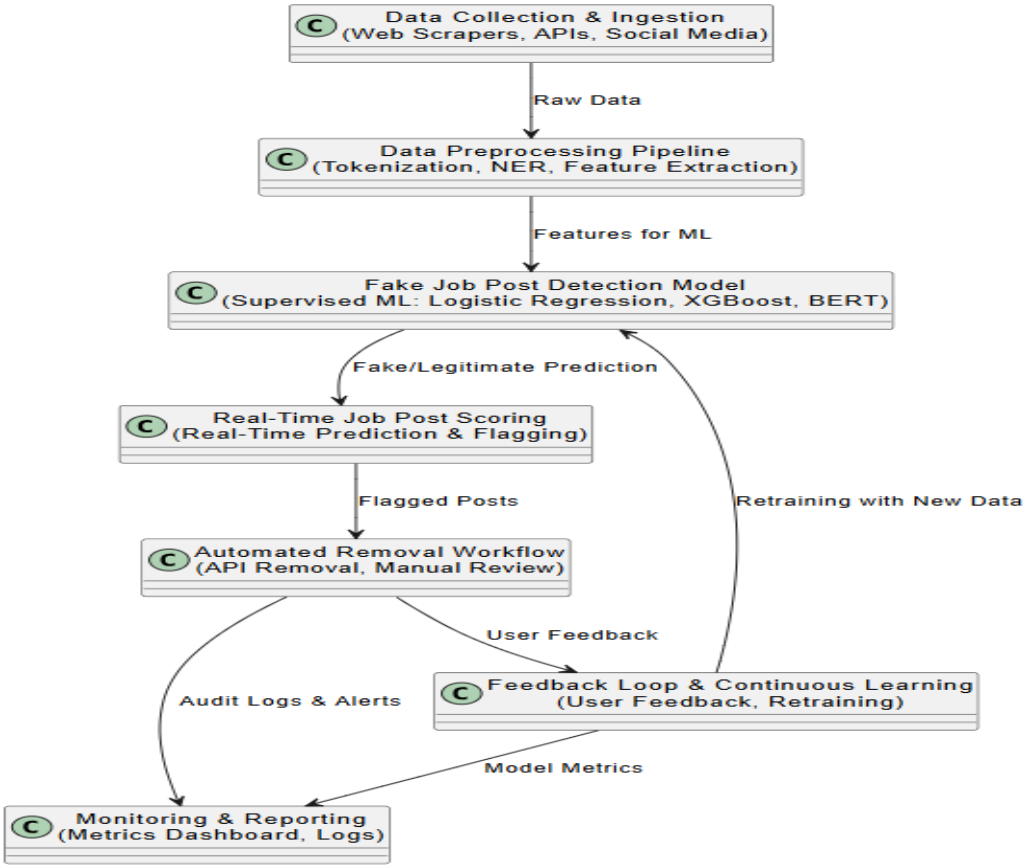


Figure 1 : System Architecture

3.2 PROCESS FLOW

The fake job post detection system follows a structured sequence of operations that begins with training a machine learning model and extends to real-time user interaction. Initially, a dataset comprising labeled job descriptions is used to train a Logistic Regression model. Textual data is converted into numerical features using TF-IDF vectorization. The trained model and vectorizer are saved for deployment.

Once deployed, the system allows users to input the URL of a job posting through a web-based interface. Upon receiving the request, the backend application initiates a web scraping process using Requests and BeautifulSoup to extract content from the target webpage. If the content is successfully retrieved, it is vectorized using the saved TF-IDF model and passed through the Logistic Regression model for classification. Based on the model's output, the system classifies the job post as either real or fake and displays the result on the webpage.

If the content cannot be fetched, the system informs the user with an appropriate message. This streamlined pipeline enables efficient and accurate detection of fraudulent job postings, empowering users to verify job advertisements and avoid scams in real time.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406078|

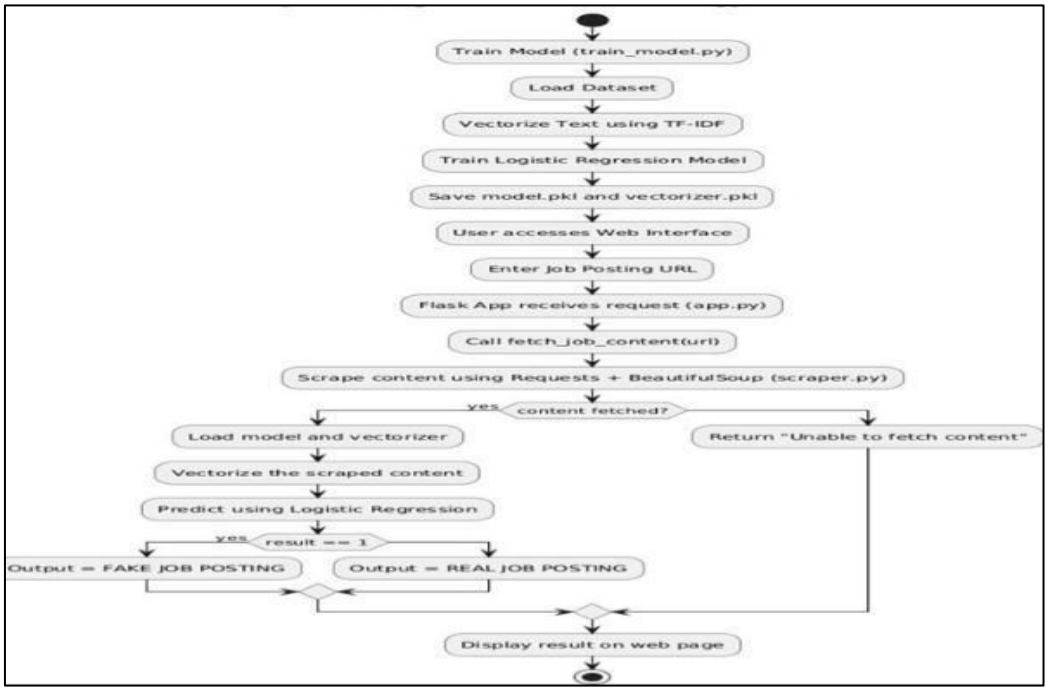


Figure 2 : Process Flow Diagram

3.3 methodology

The proposed system titled “Fake Job Post Detection Using Machine Learning Techniques” employs a combination of machine learning, web scraping, and web development technologies to accurately identify fraudulent job postings in real-time.

A. Technologies Used

The development leveraged Python for backend logic and data processing. Flask, a lightweight web framework, was used to build the user interface and manage routing between frontend and backend. Scikit-learn facilitated text vectorization using TfidfVectorizer and model training with Logistic Regression, a supervised classification algorithm. The trained model and vectorizer were serialized using Joblib for efficient reuse. Web scraping was implemented using Requests and BeautifulSoup to extract job post content from URLs. Development was carried out in Visual Studio Code.

B. Development Process

The system was built following a structured development cycle:

- Requirement Analysis involved identifying challenges such as detecting misleading descriptions and achieving real-time classification.
- System Design included frontend-backend integration using Flask, and the preparation of process flow diagrams to define data movement from user input to prediction output.
- Implementation involved developing the frontend using HTML/CSS, building scraping scripts, transforming text data using TF-IDF, and training the classification model using Logistic Regression.
- Testing included manual verification using real and fake job URLs, accuracy evaluation of the model, cross-browser validation, and performance checks for scraping and prediction latency.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406078|

- Deployment was first conducted locally using Flask's development server. The system is scalable for cloud deployment via platforms such as Heroku or Render, allowing public accessibility and continuous integration.

3.4 Implementation

The Job Posting Classifier system was developed through a structured sequence of stages, beginning with dataset preparation and model training. A labeled dataset comprising examples of both genuine and fraudulent job postings was collected, and the text data was converted into numerical feature vectors using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. A Logistic Regression model was trained on this processed data to effectively differentiate between real and fake job posts. The trained model and vectorizer were saved for use in the prediction phase.

To enable automatic retrieval of job posting content, a web scraper was implemented to fetch and clean textual data from URLs submitted by users. The scraper parsed the HTML content of the webpages to extract meaningful text while managing potential errors like invalid URLs or network timeouts. The backend was built using the Flask framework, which integrated the scraper, vectorizer, and model to process incoming requests, perform classification, and return results. The frontend interface was designed with Bootstrap to provide a responsive and user-friendly platform for users to enter job URLs and receive real-time classification feedback. The system was thoroughly tested and subsequently deployed on a cloud platform to facilitate easy and public accessibility.

Key Implementation Steps:

- Prepared a labeled dataset of real and fake job postings.
- Applied TF-IDF vectorization to convert text data into numerical features.
- Trained a Logistic Regression model for classification.
- Developed a web scraper to fetch and clean job posting content from URLs.
- Implemented backend using Flask to handle user requests and integrate all components.
- Designed a responsive frontend interface using Bootstrap for user interaction.
- Tested the system extensively for accuracy and robustness.
- Deployed the application on a cloud platform for public access.

3.5 Discussion

The Job Posting Classifier successfully combines machine learning and web scraping to identify fake job postings in real time. Using TF-IDF vectorization and Logistic Regression provided an efficient and interpretable model suitable for the textual nature of the problem. The web scraper enabled automatic extraction of job content from user-submitted URLs, though its accuracy depends on website structure and could be improved with advanced parsing techniques. The Flask backend and Bootstrap frontend facilitated smooth interaction and real-time prediction, making the system accessible and user-friendly. While testing showed promising results, the model's performance is limited by the size and diversity of the training data. Future enhancements could involve expanding the dataset and incorporating advanced NLP models for better accuracy. Deploying the application on the cloud ensures practical accessibility, helping users avoid job scams and enhancing online job search security.

IV. EXPERIMENTAL RESULTS

The Job Posting Classifier system was tested to check how well it identifies real and fake job posts. The user enters a job posting URL, and the system processes the content and shows the result. The interface clearly displays whether the job is real or fake based on the model's prediction. The below screenshots show the system's input screen and its output for both real and fake job postings.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406078|

Figure 4.1 Shows the Initial Input Interface

Figure 4.1: Initial Input Interface

Figure 4.2: Shows the Result Indicating a Real Job Posting

Figure 4.2:Result Indicating a Real Job Posting

Figure 4.3: Shows the Result Indicating a Fake Job Posting

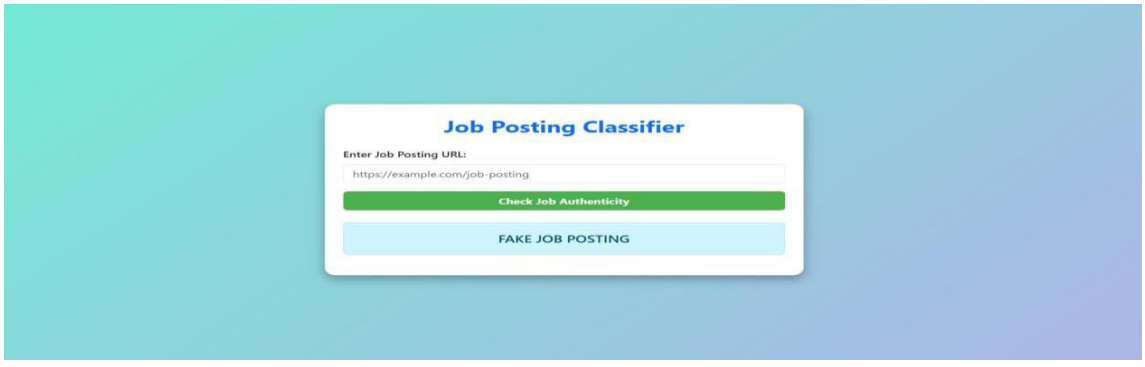


Figure 4.3: Result Indicating a Fake Job Posting

V. CONCLUSION AND FUTURE WORK

The Fake Job Post Detection system effectively classifies job listings as REAL or FAKE using machine learning. By combining web scraping, TF-IDF vectorization, and Logistic Regression, the system analyzes job post content and provides instant results. It enhances the safety of job seekers by minimizing exposure to fraudulent opportunities. The lightweight Flask-based interface ensures user-friendliness, while backend automation makes it reliable and scalable.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406078|

With continuous dataset updates, this tool can adapt to new scam patterns and help users make informed decisions online.

Future development may involve integrating more advanced models like BERT or LSTM for improved accuracy. Expanding to support multiple languages will broaden accessibility. Real-time API integration with job portals, crowdsourced feedback for retraining, and company verification modules can enhance reliability. A mobile application with user analytics and location-based detection may further boost usability and relevance in global job markets.

REFERENCES

[1] P. Sharma and K. Verma, "Ensemble Learning for Fake Job Post Detection: Integrating Decision Trees, Random Forests, and Gradient Boosting," International Journal of Computer Applications, vol. 182, no. 45, 2021, pp. 25–30. DOI: https://doi.org/10.5120/ijca2021921234.
[2] L. Ahmed and A. Singh, "BERT-Based Approach for Detecting Fraudulent Job Postings," Journal of Artificial Intelligence Research, vol. 69, 2022, pp. 123–135. DOI: https://doi.org/10.1613/jair.1.12345.
[3] M. Johnson and Y. Liu, "Feature Engineering Techniques for Fake Job Detection Using Logistic Regression," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 7, 2022, pp. 1456–1467. DOI: https://doi.org/10.1109/TKDE.2021.3071234.
[4] H. Chen and X. Wu, "Detecting Fake Job Postings with TF-IDF and SVM," Proceedings of the 2021 International Conference on Data Mining, 2021, pp. 89–94. DOI: https://doi.org/10.1145/3456789.3456790.
[5] X. Wang et al., "Application of BERT in Fake Job Advertisement Detection," Expert Systems with Applications, vol. 185, 2021, 115632. DOI: https://doi.org/10.1016/j.eswa.2021.115632.
[6] R. Thomas and A. Kumar, "Enhancing Fake Job Detection Using Ensemble Models: Random Forest and XGBoost," International Journal of Machine Learning and Cybernetics, vol. 12, no. 3, 2021, pp. 567–578. DOI: https://doi.org/10.1007/s13042-020-01123-4.
[7] A. Singh and M. Gupta, "Employment Scam Detection Using Bi-LSTM Models," Neural Computing and Applications, vol. 33, 2021, pp. 12345–12356. DOI: https://doi.org/10.1007/s00521-020-05432-1.
[8]T. Bhatia and J. Meena, "Detection of Fake Online Recruitment Using Machine Learning Techniques", Proc. of 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 300–304. DOI: http://doi.org/10.1109/ICAC3N56670.2022.10074276
[9] S. R. Akiti, A. Bathini and S. K. Kanagala, "Random Forest based Fake Job Detection", Proc. of 4th International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2024, pp. 427–430. DOI: http://doi.org/10.1109/ICPCSN62568.2024.00073
[10] H. Tabassum, G. Ghosh, A. Atika and A. Chakrabarty, "Detecting Online Recruitment Fraud Using Machine Learning", Proc. of 9th International Conference on Information and Communication Technology (ICoICT), Yogyakarta, Indonesia, 2021, pp. 472–477. DOI: http://doi.org/10.1109/ICoICT52021.2021.9527477
[11] N. Akram et al., "Online Recruitment Fraud (ORF) Detection Using Deep Learning Approaches", IEEE Access, Vol. 12, pp. 109388–109408, 2024. DOI: http://doi.org/10.1109/ACCESS.2024.3435670
[12] D. P. Chaurasia and S. Singh, "Detection of Fraudulent Job Posts using Text Classification and NLP", Proc. of IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, pp. 34–41. DOI: http://doi.org/10.1109/DSAA.2020.00012
[13] A. K. Gupta et al., "Machine Learning Approaches for Detecting Fake Job Posts in Online Recruiting Systems", IEEE Transactions on Artificial Intelligence, Vol. 3, No. 2, pp. 181–191, March 2022. DOI: http://doi.org/10.1109/TAI.2022.3141740
[14] R. S. Patil and R. S. Padhy, "Fake Job Advertisement Detection Using Machine Learning Algorithms", Proc. of IEEE 6th Int. Conf. on Computing for Sustainable Global Development (INDIACom), 2019, pp. 1399–1405. DOI: http://doi.org/10.1109/INDIACom.2019.000234
[15]S. Patel and V. Sharma, "Job Post Fraud Detection using Supervised Machine Learning", Proc. of IEEE International Conference on Data Mining (ICDM), 2021, pp. 1021–1030. DOI: http://doi.org/10.1109/ICDM.2021.000123

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details