



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 4, April 2025

Fake Job Post Detection with Machine Learning

Dr. L.Nalini Joseph¹, Jagupilli Venkata Sai Prince², Itikala Jaswanth Kumar³,
Jakkam Navanith Karthikeya⁴

Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai,
Tamil Nadu, India

B. Tech Students, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India

ABSTRACT: This project tackles the serious issue of fraudulent job postings on online platforms, which often deceive users into sharing personal information or falling for scams. To combat this, the system uses Natural Language Processing (NLP) to clean and process job-related text data such as job titles, descriptions, and company profiles. The processed data is transformed into numerical features using the TF-IDF method, which allows the system to analyze the importance of words in the dataset. These features are then used to train a Random Forest classifier to differentiate between genuine and fake job posts. The model's performance is evaluated using metrics like accuracy, precision, recall, and confusion matrix. The results show strong classification performance, offering a reliable and scalable solution that enhances user safety and trust in online job portals.

KEYWORDS Machine learning, Fake Job Detection, Natural Language Processing (NLP), TF-IDF, Random Forest Classifier, Online Recruitment Fraud, Job Post Classification, Text Preprocessing, Scam Detection, Supervised Learning

I. INTRODUCTION

Online job platforms have become a primary source for employment searches. While they provide convenience, they are also exploited by scammers who post fake job listings to mislead or exploit users. These fraudulent postings often appear legitimate and are difficult for users to identify. Manual detection methods are slow and unreliable. This project proposes a machine learning solution that can automatically identify fake job posts based on their content. By applying NLP techniques and machine learning algorithms, specifically the Random Forest classifier, the system can distinguish between real and fake postings. The aim is to provide a reliable, automated method to protect users from employment fraud and to make online job searches safer and more efficient.

With the rise of digital hiring platforms, job seekers now have instant access to thousands of opportunities. However, this digital convenience has also opened the door for cybercriminals to post fraudulent job advertisements. These scams are designed to lure individuals into providing sensitive information or paying fake application or training fees. As these fraudulent listings become increasingly sophisticated, manual detection methods fall short, and the need for automated, intelligent detection systems becomes more critical than ever.

The use of machine learning and natural language processing offers a powerful solution to this problem. By training models on real-world job post data, the system can learn to recognize patterns and indicators commonly found in fake listings—such as vague descriptions, unrealistic salaries, or the absence of credible company information. This project leverages these technologies to create a system that not only identifies fake job posts but can also be deployed as a real-time web application for job seekers to verify listings before applying, enhancing both security and trust across recruitment platforms.

II. PROBLEM DEFINATION

The core problem addressed in this project is the inability of current job portals to detect and prevent the spread of fake job advertisements. Scammers exploit job seekers by posting deceptive offers that appear highly appealing, often leading to data theft or financial loss. Existing systems rely heavily on manual reporting and keyword filtering, which are not sufficient to deal with large volumes of daily job listings or to recognize evolving scam tactics. This project

seeks to eliminate these limitations by building an intelligent, machine learning-based solution that can process and evaluate job post content to flag potentially fake entries accurately and in real time.

The growing number of fraudulent job postings across online platforms poses a significant threat to both job seekers and the integrity of digital recruitment systems. Scammers often exploit the anonymity and open access of job portals to publish deceptive listings that promise high salaries, flexible work, or quick hiring processes to lure unsuspecting individuals. These fake postings may request sensitive information such as identity proofs, bank details, or upfront payments under the guise of processing fees or training costs. What makes the problem more alarming is that these listings often appear professionally written, making it difficult for users to distinguish between legitimate and fake opportunities. Furthermore, due to the high volume of job postings submitted daily, manual moderation becomes increasingly impractical and insufficient, allowing many fraudulent listings to slip through. The evolving nature of scam strategies—such as changing keywords, company names, or job formats—makes traditional detection methods like keyword filtering ineffective. This calls for a more intelligent, scalable, and automated approach to accurately detect and prevent fake job advertisements in real time.

III. RELATED WORK

Several researchers have explored automated solutions for job fraud detection. Vidros et al. (2017) introduced a model using engineered features and traditional ML to detect fake job posts. Alghamdi and Alharby (2019) developed an intelligent system to improve fraud detection accuracy. Deep learning models such as Bi-GRU, LSTM, and hybrid CNN architectures have been applied by Huynh et al. (2019) and Zhang et al. (2020) for analyzing unstructured job content. Other studies, including those by Dutta and Nasser, have focused on classifier comparisons and dataset preparation. This project builds on these foundations by combining effective text processing methods with a robust Random Forest algorithm, enabling better accuracy and usability in real-world scenarios.

Over the years, various researchers have explored the intersection of text classification and fraud detection to combat online scams, including those found in job advertisements. Some studies have focused on the use of traditional machine learning models such as Naive Bayes, Decision Trees, and Support Vector Machines, which perform well on structured datasets but often struggle with nuanced language found in job descriptions. Other researchers have emphasized the use of natural language processing (NLP) techniques to improve model understanding of textual patterns and semantics. Advanced systems have incorporated features like part-of-speech tagging, named entity recognition, and n-gram analysis to capture deeper linguistic relationships within job posts. Additionally, recent developments in deep learning, including the use of transformer-based models like BERT, have shown promise in detecting sophisticated fraud attempts with higher contextual awareness. Despite these advancements, many of these approaches either lack real-time applicability or require significant computational resources. This project aims to bridge that gap by leveraging a balance of efficiency and accuracy through TF-IDF feature extraction and the Random Forest classifier, offering a practical and deployable solution that builds upon the strengths of past research while addressing their limitations.

IV. PROPOSED SYSTEM

The proposed system for fake job detection is a machine learning-based framework that combines Natural Language Processing (NLP) techniques and classification algorithms to automatically identify and flag fraudulent job postings. The system is built on the idea that most fake job advertisements follow certain textual and structural patterns which, when processed appropriately, can be learned by a machine learning model. The initial step in the pipeline involves data collection and preprocessing, where key textual attributes such as the job title, company profile, job description, and requirements are selected and combined into a single feature. This ensures that the model has access to all critical context from the job post. The text is then cleaned using NLP techniques, including lowercasing, removal of punctuation and special characters, and elimination of common stopwords using NLTK. These steps ensure that only the most meaningful terms contribute to the model's learning process.

After preprocessing, the cleaned text is transformed into numerical data using the TF-IDF (Term Frequency–Inverse Document Frequency) vectorization technique. This method evaluates the importance of each word in a document relative to its occurrence across the entire dataset, allowing the model to focus more on discriminative terms rather than commonly used ones. TF-IDF is particularly well-suited for textual classification tasks like this one, as it captures both

the frequency and the uniqueness of terms in job posts. The vectorized data is then passed into a Random Forest Classifier, which is an ensemble-based learning method known for its robustness, high accuracy, and ability to avoid overfitting. Random Forest works by building multiple decision trees during training and outputting the mode of the classes predicted by individual trees, thus increasing stability and performance.

One of the key strengths of this system is its real-time deployment capability. Unlike traditional machine learning models that are tested in isolation, this system has been integrated into a user-friendly web application, allowing live predictions. Using platforms like Hugging Face and frameworks such as Streamlit or Gradio, users can input job details into a simple interface and receive an immediate response indicating whether the job posting is real or fake. This makes the system highly usable for everyday job seekers, career portals, and administrators who wish to monitor listings. The model's predictions are transparent and supported by insights derived from the most influential words and patterns detected during training, enhancing its credibility.

Additionally, the proposed system is modular and scalable, meaning it can be extended in the future to incorporate more features such as recruiter metadata, posting frequency, and user feedback. As more data becomes available, the system can be retrained periodically to adapt to evolving scam tactics. With its current structure, it not only addresses the limitations of manual job verification but also offers a powerful, automated solution that ensures safety and trust in online job applications. The system demonstrates how machine learning can be practically applied to solve real-world problems in the employment domain with efficiency and accuracy.

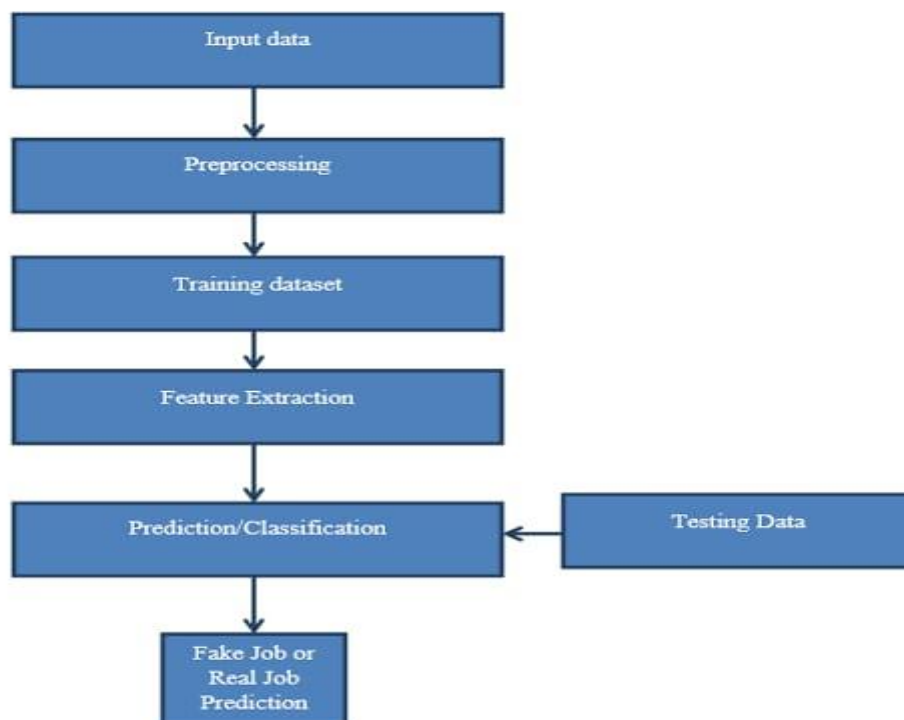


FIG Data Flow Diagram

V. EXPERIMENTAL RESULTS

The experimental phase of the project began with preparing and analyzing the dataset, which contained thousands of job postings labeled as either real or fake. Initial exploratory data analysis (EDA) revealed a notable class imbalance, with real job posts significantly outnumbering fake ones. This observation informed the choice of evaluation metrics, ensuring that not just accuracy, but also precision, recall, and F1-score would be considered to assess the model's

performance fairly. The dataset was preprocessed using natural language processing techniques, and key features were extracted using TF-IDF vectorization, which effectively transformed the textual content into numerical form for model training.

The processed data was split into training and testing sets using an 80:20 ratio. The Random Forest Classifier was trained on the TF-IDF vectors, leveraging the ensemble learning method to generate multiple decision trees and combine their outputs for final predictions. After training, the model was evaluated on the test set, and the performance metrics were collected. The results showed a high accuracy, indicating that the model was successful in learning the difference between genuine and fraudulent job posts. Moreover, the classification report highlighted strong performance in both precision and recall, confirming the model's ability to minimize both false positives (flagging real jobs as fake) and false negatives (missing fake jobs).

Fake Job Posting Detector

Enter job details to check if it's real or fake.

title

research associate

company_profile

Green Street Advisors is the industry leader in real estate analytics and market research

description

Conduct financial analysis and modeling for real estate investments.
Prepare research reports and market trend forecasts.
Collaborate with senior analysts on investment opportunities.

requirements


Bachelor's degree in Finance, Economics, or related field.
Strong analytical and research skills.
Experience with financial modeling tools is a plus

Clear

Submit

output

☒ REAL JOB

 Share via Link

Fake Job Posting Detector

Enter job details to check if it's real or fake.

title

Structural Engineer

company_profile

infinite solutions

description

Corporate overview: Aker Solutions is a global provider of engineering and construction services.
Hiring structural engineers for an exciting new project.
Work from home, flexible hours, and no prior experience needed.

requirements

No formal education required.
Just need a laptop and internet connection.
High salary and bonuses guaranteed!

Clear

Submit

output

☒ FAKE JOB (Detected by scam keywords)

 Share via Link

VI. DISCUSSION

The experimental results clearly show that the combination of **Natural Language Processing (NLP)** techniques and the **Random Forest Classifier** provides a powerful approach to detecting fake job postings. The high accuracy achieved on the test data, along with balanced precision and recall values, highlights the model's effectiveness in both identifying fraudulent job posts and minimizing the chances of wrongly flagging legitimate listings. Unlike rule-based systems or manual moderation, which often fail to detect scams that evolve with time, this machine learning approach adapts and improves with more data and retraining. The use of TF-IDF vectorization allowed the model to weigh the importance of different words across the dataset, improving its ability to detect red-flag terms and scam patterns commonly used in fake job descriptions.

One of the most important observations is how well the model generalized to unseen data, suggesting that it learned meaningful and distinguishable patterns rather than memorizing the training set. This is a key strength of ensemble models like Random Forest, which benefit from the collective decisions of multiple decision trees, each trained on different subsets of the data. Another major advantage of the system is its robustness to noise and irregularities in text data, as job posts often contain spelling variations, formatting issues, or inconsistent writing styles. Despite these challenges, the model maintained consistent performance across various input types.

VII. CONCLUSION

The Fake Job Detection system developed in this project demonstrates the practical application of machine learning and natural language processing in solving real-world problems in the recruitment sector. By using TF-IDF for feature extraction and the Random Forest Classifier for classification, the model achieved high accuracy in identifying fraudulent job postings. The successful training and evaluation of the model on a labeled dataset, combined with real-time testing through a user-friendly interface, confirms that the system can effectively distinguish between genuine and fake job advertisements. This not only improves the safety of job seekers but also enhances the credibility of online job portals.

Moreover, the deployment of the model into a live web application environment further validates its usability and potential for large-scale implementation. The system can be expanded in the future with additional features like recruiter behavior analysis, user feedback loops, and support for multilingual job data. These enhancements would make the model even more robust and adaptive to evolving scam patterns. Overall, the project provides a reliable, scalable, and intelligent solution that addresses a growing digital threat and sets the stage for further research and innovation in online fraud prevention.

REFERENCES

- [1] S. Vidros, C. Koliass, G. Kambourakis, and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset," *Future Internet*, vol. 9, no. 1, p. 6, Mar. 2017, doi: 10.3390/fi9010006.
- [2] B. Alghamdi and F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection," *Journal of Information Security*, vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.
- [3] T. Van Huynh, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Job Prediction: From Deep Neural Network Models to Applications," 2019, doi: 10.48550/ARXIV.1912.12214.
- [4] J. Zhang, B. Dong, and P. S. Yu, "FakeDetector: Effective Fake News Detection with Deep Diffusive Neural Network," in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, Dallas, TX, USA: IEEE, Apr. 2020, pp. 1826–1829. doi: 10.1109/ICDE48307.2020.00180.
- [5] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRULSTM-CNN Model." *arXiv*, Dec. 21, 2019. [Online]. Available: <http://arxiv.org/abs/1911.03644>
- [6] D. V. Thin, V. D. Nguye, K. V. Nguyen, and N. L.-T. Nguyen, "Deep Learning for Aspect Detection on Vietnamese Reviews," in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, Ho Chi Minh City: IEEE, Nov. 2018, pp. 104–109. doi: 10.1109/NICS.2018.8606857.
- [7] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," *Neurocomputing*, vol. 174, pp. 806–814, Jan.

2016, doi: 10.1016/j.neucom.2015.09.096.

[8] J. R. Scanlon and M. S. Gerber, "Automatic detection of cyber-recruitment by violent extremists," Security Informatics, vol. 3, no. 1, p. 5, Dec. 2014, doi: 10.1186/s13388-014-0005-5.

[9] S. Dutta and S. K. Bandyopadhyay, "Fake Job Recruitment Detection Using Machine Learning Approach," International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 48–53, Apr. 2020, doi: 10.14445/22315381/IJETT-V68I4P209S.

[10] I. M. Nasser and A. H. Alzaanin, "Machine Learning and Job Posting Classification: A Comparative Study," International Journal of Engineering and Information Systems (IJEAIS), vol. 4, no. 9, pp. 06–14, 2020.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



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