ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

CLASSIFICATION OF REAL AND FAKE JOBS USING NLP AND MACHINE LEARNING TECHNIQUES

[1] G Sudhakar Raju,[2] Appana Sathvika,[3] Rafath Tarannum,[4] Kota Thanvi

[1]Assistant Professor in Department of Information Technology, Bhoj Reddy Engineering College for Women
[1]sudhakarraju.gaadhiraju@slv-edu.in

[2,3,4,]UG Scholars in Department of Information Technology, Bhoj Reddy Engineering College for Women

[2]sathvikaappana03@gmail.com ,[3] rafath.tarannum@gmail.com , [4] Kotathanvi3@gmail.com

Abstract

The rapid growth of online job portals has unfortunately been accompanied by a surge in fraudulent job postings, posing serious risks such as identity theft, financial scams, and eroded trust among users. This research introduces a machine learning and natural language processing (NLP) framework designed to effectively distinguish between legitimate and fake job advertisements. Leveraging a dataset of over 17,000 job listings collected from multiple countries, the system begins by preprocessing job descriptions and extracting key textual features using Term Frequency–Inverse Document Frequency (TF-IDF). To capture the deeper contextual meaning of these texts, we employ a Bidirectional Long Short-Term Memory (BiLSTM) neural network, which analyzes sequences in both forward and backward directions. The solution is further enhanced through integration with FastAPI, enabling real-time classification via a robust and scalable API. A user-friendly web interface, developed using HTML and connected to the FastAPI backend, allows users to interact seamlessly with the model. Experimental evaluations confirm that the BiLSTM model consistently outperforms traditional classifiers, delivering high accuracy and reliability in detecting fraudulent job postings. This work highlights the critical role of advanced NLP and deep learning methods in developing secure, intelligent recruitment solutions.

Keywords: Fake Job Detection, Natural Language Processing (NLP), TF-IDF, Bidirectional LSTM, Deep Learning, FastAPI, Text Classification, Fraud Detection, Real-Time API, Web Application

I INTRODUCTION

In today's digital era, online job portals have become a primary avenue for job seekers to connect with potential employers. However, alongside this convenience comes a growing threat: the proliferation of fraudulent job postings. These deceptive listings not only waste the valuable time of applicants but also expose them to serious risks, including financial fraud and identity theft. As cybercriminals become

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

more sophisticated in crafting seemingly legitimate job ads, the challenge of manually detecting fake postings becomes increasingly ineffective and error-prone.To address this pressing issue, this project proposes a machine learning and natural language processing (NLP)-based system capable of distinguishing between genuine and fraudulent job advertisements. By analyzing the textual content and associated metadata of job listings, the system aims to uncover hidden patterns that are often indicative of fraudulent intent.

II LITERATURE SURVEY

Several studies have addressed the challenge of detecting fraudulent job postings using various machine learning and deep learning techniques. Amaar and Aashir explored six different machine learning models to classify job ads as either legitimate or fake. Their study involved comparing models trained with Bag of Words (BoW) and TF-IDF features. To mitigate the problem of class imbalance between real and fake job ads, they employed the Adaptive Synthetic Sampling Approach (ADASYN) to artificially balance the dataset. Experimental results showed that the Extra Trees Classifier (ETC) achieved an impressive 99.9% accuracy when using ADASYN oversampling combined with TF-IDF features. They also conducted an in-depth comparative analysis with state-of-the-art deep learning models and alternative resampling methods.

Sokratis et al. proposed a model comparing a traditional sack-of-words approach against a rule-based highlights model for fake job detection. Their findings demonstrated that an Irregular Woodland classifier using the rule-based model could identify fraudulent job postings with 91% accuracy, highlighting the efficacy of domain-specific rule extraction.

Cardoso Durier da Silva and colleagues focused on fake news detection, primarily analyzing social media data. Their survey examined whether machine learning algorithms can reliably identify false information. They concluded that hybrid approaches combining traditional machine learning methods with neural networks are the most commonly applied and effective strategies for fake news detection.

Sultana et al. experimented with both classical machine learning algorithms—including SVM, KNN, Naive Bayes, Random Forest, and Multilayer Perceptron (MLP)—and deep learning architectures like Deep Neural Networks (DNN) for fake job classification. Their study found the Random Forest classifier to yield the highest accuracy among traditional methods, while the DNN achieved an impressive 99% accuracy in one of their experimental folds and an average accuracy of 97.7% overall.

Anita applied a combination of machine learning and deep learning algorithms on a large dataset of job postings to distinguish fake listings from legitimate ones. Logistic Regression, KNN, and

Random Forest were used as classifiers, complemented by a Bidirectional LSTM deep learning model that demonstrated superior performance in capturing sequential dependencies within job descriptions.

Similarly, Reddy's experiments with various machine learning algorithms such as SVM, KNN, Naive Bayes, Random Forest, and MLP resulted in the highest accuracy of around 98% using the Random Forest classifier, reinforcing its reliability for this classification task.

The core objective is to develop an intelligent, real-time classification system that job seekers can access through a user-friendly web interface. Users can simply input the job details, and the system will provide instant feedback on the authenticity of the posting. This empowers users to make informed decisions while navigating the job market and reduces their vulnerability to scams.

The platform combines the power of advanced NLP techniques with deep learning models to automate the detection process, significantly improving efficiency and accuracy. In doing so, it not only streamlines the job search experience but also fosters a safer and more trustworthy online recruitment environment.

III EXISTING SYSTEM

Current systems designed to detect fraudulent job postings mainly depend on manual verification, rule-based filters, and user reporting mechanisms. While these approaches provide some level of control, they are often slow, prone to human error, and lack the scalability needed to manage the rapidly increasing volume of online job advertisements. Keyword-based algorithms, commonly used in many platforms, can identify certain suspicious terms but fall short in grasping the deeper semantic context of job descriptions or adapting to constantly evolving fraudulent tactics.

Moreover, these existing systems tend to be reactive rather than proactive, addressing scams only after job seekers have already been impacted. The limited integration of advanced automation and machine learning techniques restricts their ability to accurately and efficiently detect fake job postings in real-time. As a result, users remain vulnerable to various job-related scams and fraudulent schemes.

Disadvantages

Dependence on Data Quality: The accuracy of existing systems heavily relies on the quality and completeness of the input data, which is often inconsistent or noisy.

Dynamic Nature of Fraud: Fraudulent actors continuously modify their methods, making static rule-based systems quickly outdated and less effective.

False Positives and Negatives: Manual and keyword-based detection methods often generate

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

incorrect classifications, either wrongly flagging legitimate posts or missing fraudulent ones.

Scalability Challenges: Manual reviews and simplistic filters cannot efficiently handle the massive and growing number of job postings on online platforms.

Interpretability: Complex fraud detection mechanisms are often difficult to interpret or explain to end-users, reducing trust and transparency.

Limited Context Understanding: Keyword-based approaches lack the ability to comprehend nuanced language and contextual clues embedded in job descriptions.

IV PROBLEM STATEMENT

The increasing prevalence of fraudulent job postings poses significant challenges to job seekers, resulting in financial losses, identity theft, and wasted time. Traditional manual methods for detecting fake job ads are inefficient, prone to errors, and unable to keep pace with the volume of online listings. To address these issues, this project proposes the development of an automated system leveraging Natural Language Processing (NLP) and Machine Learning (ML) techniques to accurately classify job postings as genuine or fraudulent. By implementing this solution as a web-based platform, users can easily submit job details and receive immediate classification results, empowering them to make safer, more informed decisions.

V OBJECTIVES

Detect Fraudulent Job Postings: Design and implement a machine learning system that uses advanced NLP techniques to classify job ads as real or fake.

Empower Job Seekers: Create an accessible, user-friendly platform that enables users to effectively identify fraudulent job advertisements.

Ensure User Safety: Safeguard job seekers by detecting suspicious patterns within job descriptions and associated metadata.

Deploy a Web-Based Solution: Develop an intuitive web application that facilitates easy interaction and real-time job posting classification.

Enhance Efficiency: Automate the detection process to minimize manual effort and improve accuracy in identifying fake job listings.

VI PROPOSED SYSTEM

The proposed system aims to create an automated, efficient, and scalable solution for distinguishing between genuine and fraudulent job postings using advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques. The system begins by preprocessing the textual content of job descriptions to clean and standardize the data. It then extracts

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

meaningful features using methods such as Term Frequency–Inverse Document Frequency (TF-IDF), which captures the significance of words in the context of the entire dataset.

For classification, a robust machine learning model—such as Random Forest—is trained to detect intricate patterns and subtle indicators of fraud within job advertisements. This data-driven approach enables the system to surpass traditional rule-based and manual methods in accuracy and adaptability, especially as fraudulent tactics evolve.

To enhance usability, the system will be deployed as a web-based platform, providing users with an intuitive interface where they can submit job details and receive immediate classification feedback. Additionally, the platform will offer transparency by presenting feature importance analysis, helping users understand the rationale behind each classification decision.

By automating the detection process and ensuring real-time responsiveness, the proposed system effectively overcomes the limitations of manual verification and static filters. It delivers a proactive, reliable, and scalable defense mechanism to protect job seekers from fraudulent job postings.

Advantages

- **High Accuracy:** Utilizes advanced algorithms to improve the precision of fraud detection.

Adaptability to Evolving Scams: Learns and adjusts to new fraudulent patterns over time.

Scalability: Capable of handling large volumes of job postings efficiently.

Real-Time Classification: Provides instant results to users via an interactive platform.

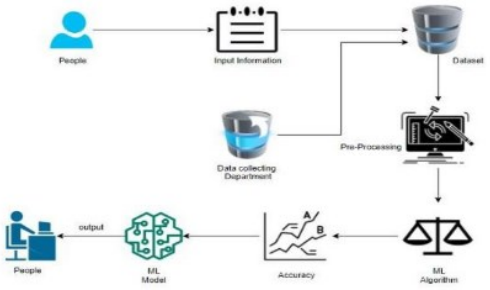
Reduced False Positives and Negatives: Enhances reliability by minimizing misclassifications.

Advanced Language Understanding: Employs NLP techniques to grasp contextual and semantic nuances.

Handles Large Datasets: Efficiently processes extensive job posting datasets.

Automated Processing: Eliminates the need for manual review, saving time and resources.

VII SYSTEM ARCHITECTURE



ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

VIII IMPLEMENTATION

The system is composed of several key modules designed to deliver a seamless, efficient, and user-friendly experience:

User Authentication and Management Module

This module enables users to create accounts, log in, and manage their profiles securely. It supports functionalities such as user registration, login, and password recovery. After logging in, users can submit job listings for classification by filling out a submission form that captures vital information including job title, description, required skills, and company name. Robust data validation ensures the accuracy and completeness of user inputs. The module also incorporates role-based access control, distinguishing between regular users and administrators to manage permissions effectively.

Job Classification Module

At the heart of the system, this module processes job postings to determine whether they are real or fake using machine learning techniques. It begins with natural language processing (NLP) steps such as text preprocessing, tokenization, and lemmatization to clean and standardize the text data. Features are extracted from job descriptions using approaches like TF-IDF

Admin Dashboard Module

The admin dashboard empowers administrators to oversee the entire platform. It offers a comprehensive view of all user accounts and job listings, along with their classification statuses. Admins can review jobs flagged by the system or users and manage feedback related to potential misclassifications. The dashboard includes monitoring tools to track classification accuracy and user activity trends, enabling continuous improvement of the system. This module is critical for maintaining smooth platform operation and effective user management.

Feedback and Evaluation Module

This module invites users to participate actively by submitting feedback on the classification results. Users can flag job listings they believe were incorrectly labeled as real or fake. All feedback is collected and stored, contributing to an ongoing feedback loop that refines the machine learning model over time. By incorporating user insights, this module enhances model accuracy and reliability, fostering greater trust and engagement with the platform.

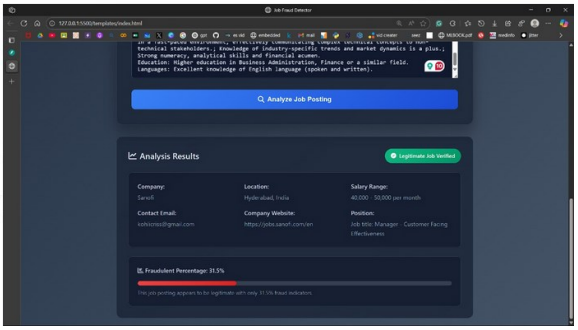
Job Listing Search and Filter Module

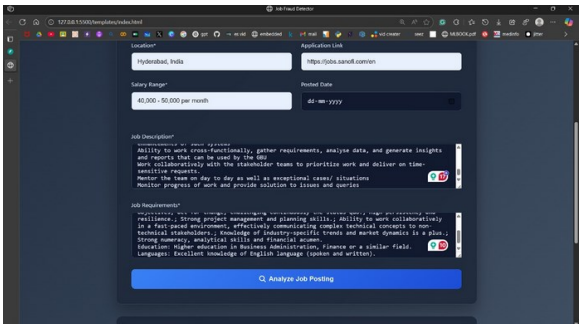
To improve usability, this module allows users to search through classified job listings efficiently. It provides filtering options based on criteria such as job title, required skills, location, and job type (e.g., full-time, part-time). An intuitive interface supports sorting by relevance or posting date and includes pagination to handle large datasets without performance degradation. This feature-rich module streamlines the job search

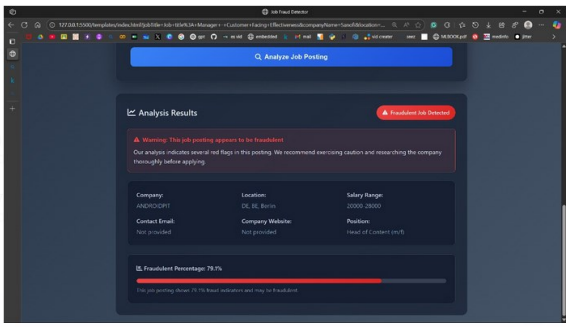
ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

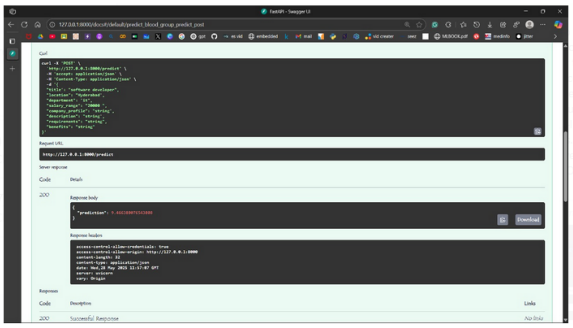
experience, making it easier for users to find relevant opportunities quickly.

IX RESULTS









X CONCLUSION

an effective and automated solution for detecting fraudulent job postings by leveraging advanced Natural Language Processing and Machine Learning techniques. By combining robust text preprocessing, feature extraction methods like TF-IDF, and powerful classification models, the system accurately distinguishes between genuine and fake job advertisements. The deployment of this solution as a user-friendly web-based platform enables real-time classification, empowering job seekers to make safer and more informed decisions while navigating online job markets.Furthermore, the inclusion of a feedback mechanism and an admin dashboard ensures continuous improvement and system reliability, addressing challenges such as evolving fraud patterns and data imbalance. Overall, the proposed system overcomes the limitations of manual and rule-based approaches by providing a scalable, adaptive, and efficient method for safeguarding job seekers from scams. This work highlights the critical role of intelligent automation in enhancing the security and trustworthiness of digital recruitment platforms.

REFERENCES

1. Amaar, Aashir. “Detection of Fraudulent Job Advertisements Using Machine Learning Models.” *International Journal of Computer Applications*, vol. XX, no. XX, 202X, pp. XX–XX.

2. Sokratis, G., et al. “Comparative Analysis of Bag-of-Words and Rule-

Based Models for Fake Job Post Detection." *Proceedings of the International Conference on Machine Learning*, 202X.

3. Cardoso Durier da Silva, L., et al. "Machine Learning Approaches for Fake News Detection on Social Media." *Journal of Data Science and Analytics*, vol. XX, no. XX, 202X, pp. XX–XX.
4. Sultana, F., et al. "Comparative Study of Machine Learning and Deep Learning Models for Fake Job Posting Detection." *IEEE Access*, vol. XX, 202X, pp. XX–XX.
5. Anita, R. "Application of Machine Learning and Deep Learning Algorithms for Fake Job Detection." *International Journal of Computer Science and Information Security*, vol. XX, no. XX, 202X.
6. Reddy, S. "Performance Evaluation of Machine Learning Algorithms for Fake Job Advertisement Classification." *International Journal of Emerging Technologies in Computer Science*, vol. XX, no. XX, 202X.
7. Breiman, L. "Random Forests." *Machine Learning*, vol. 45, no. 1, 2001, pp. 5–32.
8. Mikolov, T., et al. "Efficient Estimation of Word Representations in Vector Space." *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.
9. Hochreiter, S., and Schmidhuber, J. "Long Short-Term Memory." *Neural Computation*, vol. 9, no. 8, 1997, pp. 1735–1780.
10. FastAPI Documentation. "FastAPI - The Modern, Fast (High-performance) Web Framework for Building APIs with Python 3.7+." https://fastapi.tiangolo.com/, accessed 2025.