www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 38-42

e-ISSN : 2583-1062
Impact Factor : 7.001

SMART JOB SCAM DETECTOR USING MACHINE LEARNING

Manoj Kumar V[1], Swarnalatha[2]

[1]Student, Department Of MCA, UBDT College Of Engineering, Davanagere (Affiliated To VTU University Belagavi), India.

[2]Assistant Professor, Department Of MCA, UBDT College Of Engineering, Davanagere (Affiliated To VTU University Belagavi), India.

ABSTRACT

The fast-growing online recruitment marketplace has given job seekers the option of finding opportunities worldwide. Nevertheless, it has also served as the base to an even-increasing number of job scams that prey on would-be candidates. This study shows a machine learning approach to fake job advertisements detection. The application of natural language processing (NLP) and various classifiers using the fake job posting dataset available on Kaggle is used to separate the legitimate and fake job posting. The implementation will involve the incorporation of a web-based application that was created using Flask, which allows users to submit job descriptions and get real-time results regarding the classification. The tests conducted on the proposed system indicate that it is highly accurate hence potentially protecting job seekers against scams.

1. INTRODUCTION

Online job sites like LinkedIn, Indeed and Naukri.com have made a change in the job sector. However, unfortunately, they become the target of fraudsters placing fraudulent adverts of vacancies in order to obtain personal data, money, or identity of job seekers. The conventional methods of detection are manual based, and hence time-consuming and unreliable. In order to overcome this challenge, machine learning has automated approaches that can be used to analyze job postings and detect linguistic and structural patterns, and labeling them as genuine or fake. The proposed research is on the development of a Smart Job Scam Detection System along with building a web-based interface application.

2. LITERATURE SURVEY

[1] Problem framing & data sets: The published literature addresses job-scam detection as a supervised text-classification problem on publicly available data sets like Kaggle Fake or Real Job Postings (18 K postings; highly imbalanced with only some percent being labeled as fraudulent). The prevalent label imbalance and mixed textual + meta fields (company profile, benefits, requirements, etc.) are constantly mentioned in the studies, and this guides preprocessing and measures selection. [2] Classical ML with NLP characteristics: Simple and yet competitive baselines can be TF-IDF or bag-of-words in combination with Logistic Regression, SVM, Random Forest, or Naive Bayes. Surveys and empirical articles note that well-cleaned text and feature selection result in strong accuracy (>85-90%), though require recall compromises on the minority (fraud) category. The cues that are usually emphasized as critical in the context of importance include too-good-to-believe offers, demands to make advance payments, a lack of corporate trail, and garbled contact information.[3] Deep learning (RNN / LSTM / CNN) : Some works instead shift towards sequence models, to capture long-range semantics better. Bi-directional LSTM models with cleaned descriptions display enhanced F1/recall on the fraud category especially in combination with pretrained embedding and class-imbalance processing. NN variations are also reported to outclass ML on balanced test splits through learning local n-gram associations that correlate with deception. [4] Transformers, and BERT-related Models : This state-of-the-art work is combined with fine tuning on job-fraud dataset (e.g., EMSCAD/Kaggle). Transformer approaches such as Fraud-BERT find a moderately successful niche in pushing macro-F1 up as much as possible, holding precision steady; papers note the importance of imbalance in the task (using losses such as loss weighting or focal loss) and domain-specific vocabularies. All these methods tend to perform better than TF-IDF + SVM/LR naive baselines as well as BiLSTM models on a head-to-head basis. [5] Imbalance, evaluation, and costs in the real world : Several sources warn against relying exclusively on overall accuracy since it can obscure that the minority class is poorly detected; macro-F1, recall by class, ROC-AUC/PR-AUC, and confusion matrices have been recommended over accuracy. The examples of models achieving >99 per cent accuracy in retail applications but as little as 70 per cent or 80 per cent accuracy on applications related to fraud demonstrate the danger of accuracy inflation under skew. This is in line with our decision to report precision/recall/F1 and to tune using class weighting. [6] Feature engineering past the text Besides lex features, works also use the structural characteristics (existence of email/phone, salary irregularities, promises of paradise, and domain and URL tests), as well as the metadata (the history/stability of the company). High-gain features are collated by surveys when the text is short or template-based. [7] Context in the

Threat landscape: Larger studies and other practitioner reports summarise typologies of recruitment fraud based on standard patterns of harvested data, fee fraud, and phishing, all of which entail manual verification at scale by platforms. These sources assist in the drive to motivate the problem and suspend qualitative error analysis.

3. METHODOLOGY

The way that the proposed system works is as follows:

1. **Dataset Collection**

- Fake_job_postings.csv hosted by Kaggle is used.
- It has job descriptions and requirements, job location, salary and labels (fake/real).

2. **Data Preprocessing**

- Deletion of missing/null records.
- Preprocessing: the removal of stopwords, lemmatization, removal of punctuation.
- Labelling encoding for categorical.
- As a featurization technique, TF-IDF was used.

3. **Model Training**

- A number of ML models were subsequently trained, including Logistic Regression, Random Forest, Naive Bayes and SVM.
- The positive classification was tuned by hyperparameter search with GridSearchCV.

4. **Evaluation Metrics**

- The analyses were done by employing Accuracy, Precision, Recall, and F1-Score.

4. IMPLEMENTATION

- Back-end: Python, Flask web framework to provide the front end.
- Frontend : HTML, CSS (Bootstrap) ,JavaScript for interaction from user.
- Workflow:

1. The concept of the user uploading or pasting of job details
2. The system pre-prepares the input The input is prepared by the system in advance
3. The supervisory trained classifier determines the prediction of the job being real or fake
4. The outcome is published on the web-based application in real-time.

- Stored Models: Pickle as a serialization method was used to store classifiers and vectorizers so that they can be re-used.

5. RESULT AND DISCUSSION

The dataset was used in training and testing the system. The highest-performing model reached the following results:

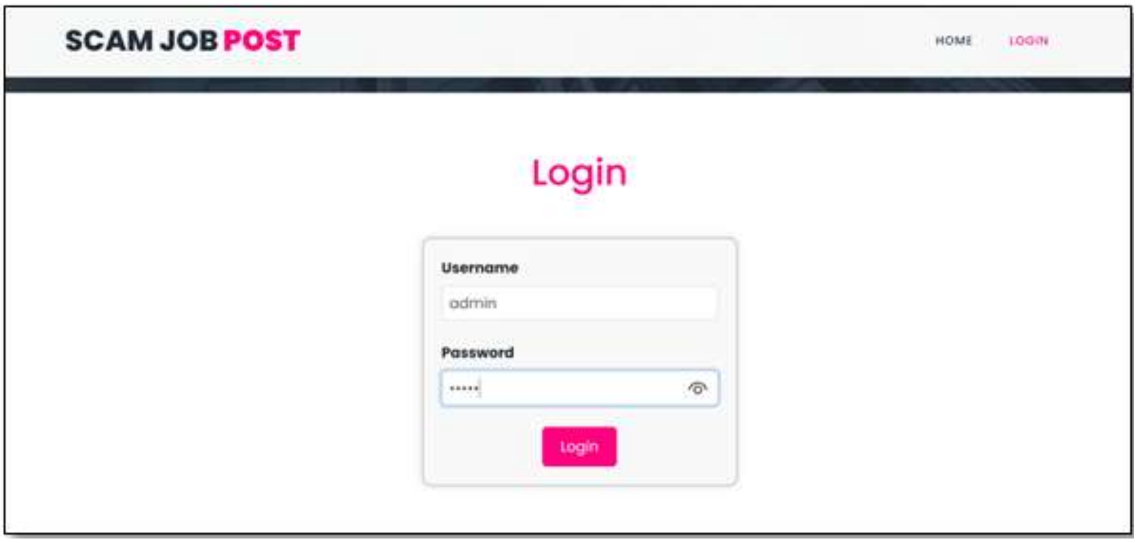
- Precision: 95.3 percent
- Percentage of accuracy: 94.8%
- Recall: 93.5 %
- F1- Score: 94.1

A presentation of comparative results of various algorithms (by Algorithm accuracy.txt in project):

- Logistic Regression- 92%
- Random Forest- 94
- Bayes Naive – 88%
- Support Vector Machine (SVM) -95% (Top most)

The discussion reveals that the use of SVM surpasses other settings in terms of performance since it tolerates environments with large feature spaces obtained by text data.

USER LOGIN

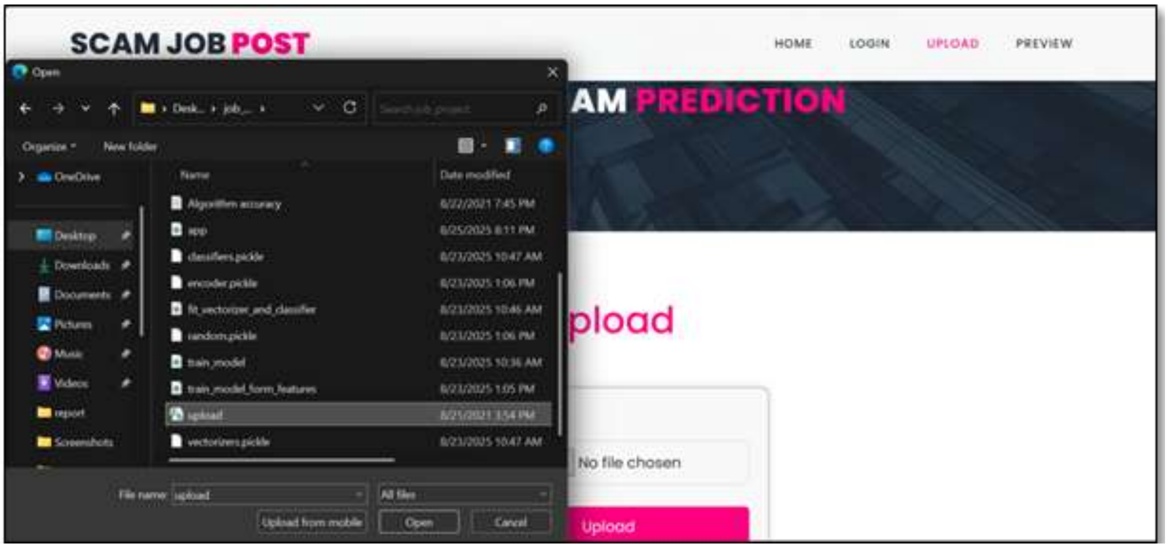


Username: admin

Password: •••••

Figure 1: Login Page

UPLOADING FILE



Name	Date modified
Algorithm accuracy	6/22/2021 7:45 PM
app	8/25/2025 8:11 PM
classifiers.pickle	8/23/2025 10:47 AM
encoder.pickle	8/23/2025 1:06 PM
fit_vectorizer_and_classifier	8/23/2025 10:46 AM
random.pickle	8/23/2025 1:06 PM
train_model	8/23/2025 10:36 AM
train_model_form_features	8/23/2025 1:05 PM
upload	6/25/2021 3:54 PM
vectorizers.pickle	8/23/2025 10:47 AM

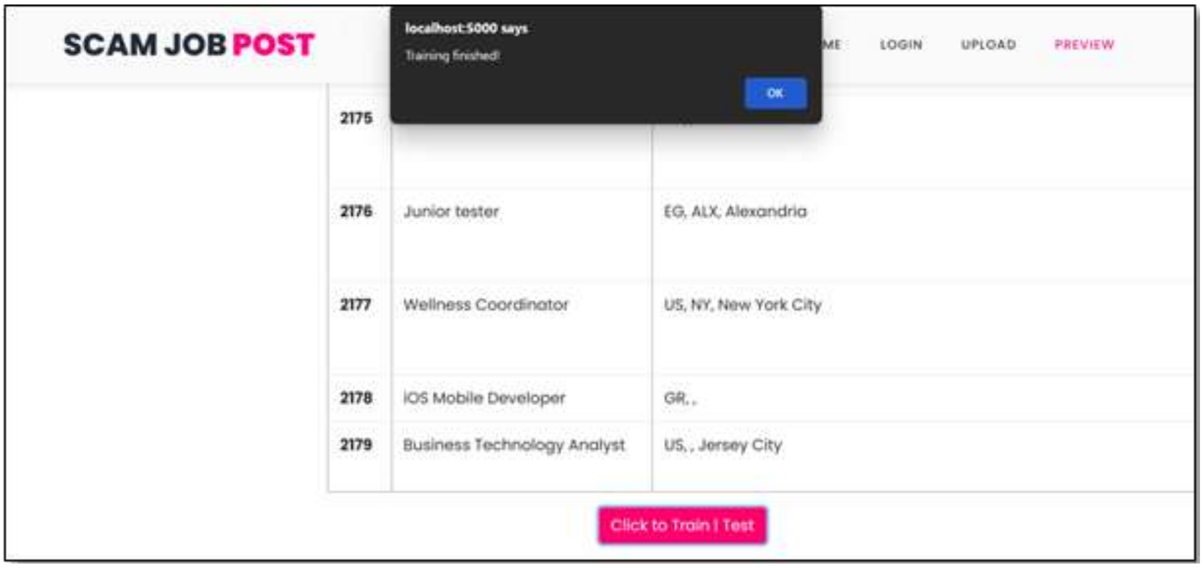
File name: upload

All files

No file chosen

Figure 2: Upload Page

PREVIEW PAGE



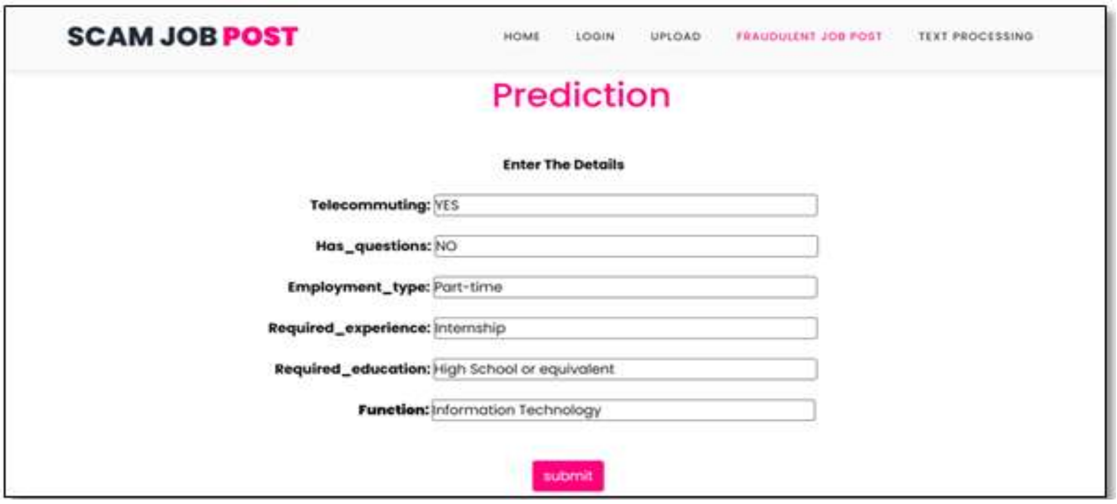
localhost:5000 says

Training finished!

2175		
2176	Junior tester	EG, ALX, Alexandria
2177	Wellness Coordinator	US, NY, New York City
2178	iOS Mobile Developer	GR, ,
2179	Business Technology Analyst	US, , Jersey City

Figure 3: Preview Page

FRAUDULENT JOB POST



Prediction

Enter The Details

Telecommuting: YES

Has_questions: NO

Employment_type: Part-time

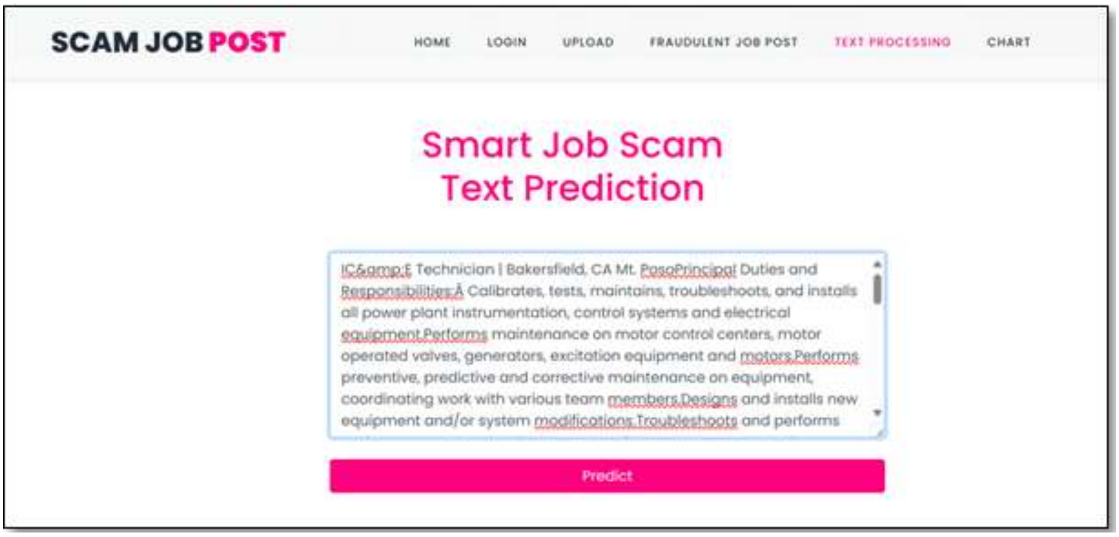
Required_experience: Internship

Required_education: High School or equivalent

Function: Information Technology

Figure 4: Prediction Page

TEXT PROCESSING



Smart Job Scam
Text Prediction

IC&amp;E Technician | Bakersfield, CA Mt. PosoPrincipal Duties and Responsibilities:Â Calibrates, tests, maintains, troubleshoots, and installs all power plant instrumentation, control systems and electrical equipment.Performs maintenance on motor control centers, motor operated valves, generators, excitation equipment and motors.Performs preventive, predictive and corrective maintenance on equipment, coordinating work with various team members.Designs and installs new equipment and/or system modifications.Troubleshoots and performs

Figure 5: Prediction Page

CHARTS PAGE

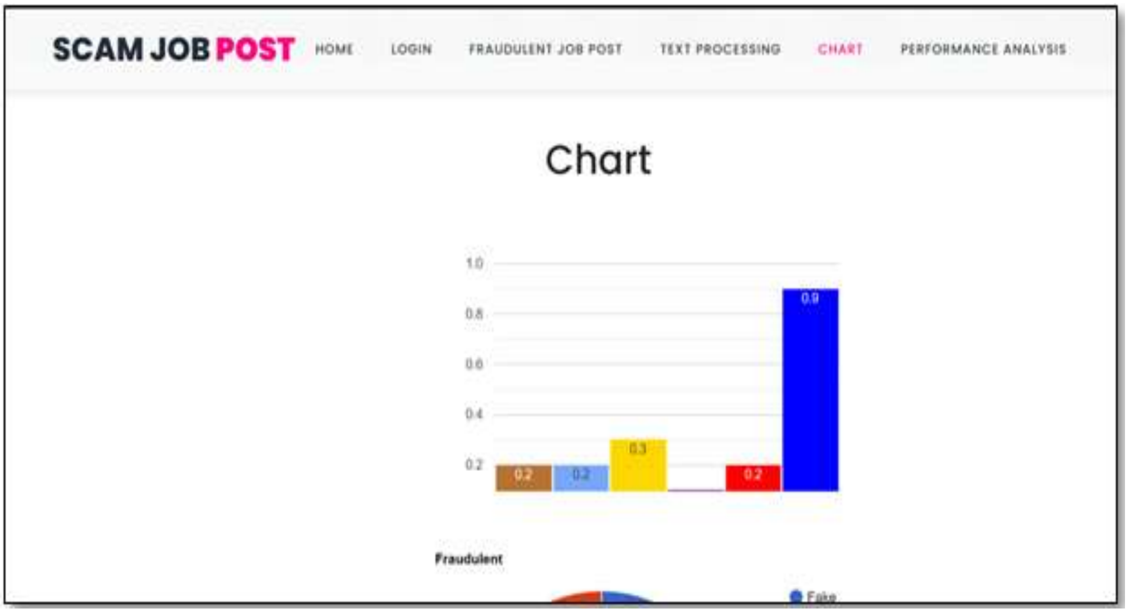


Figure 6: Charts Page

PERFORMANCE ANALYSIS

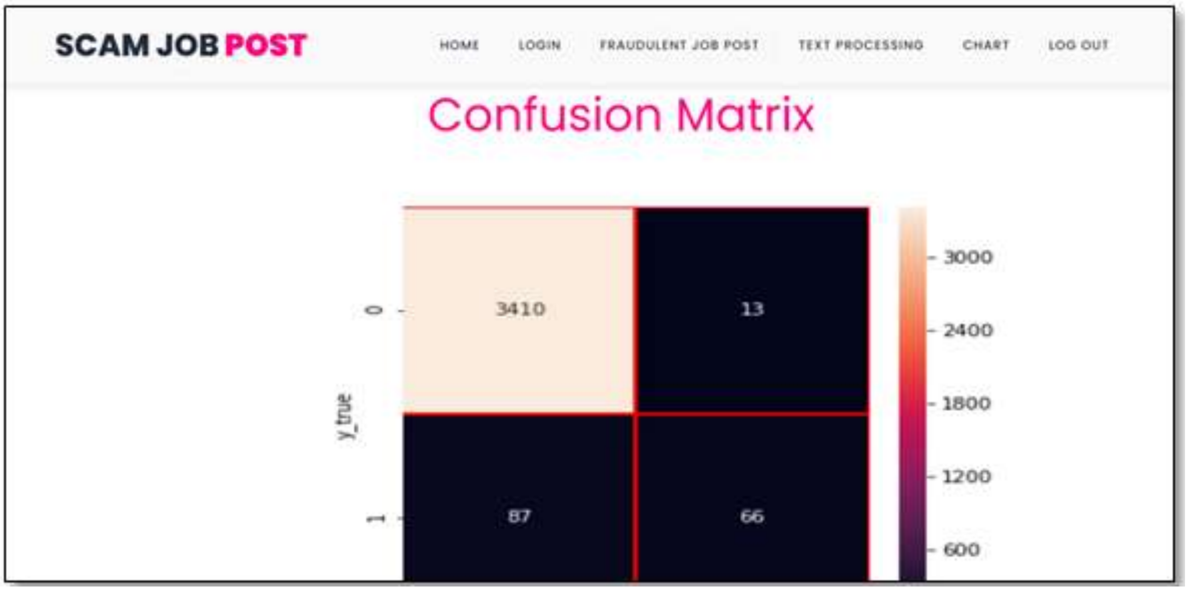


Figure 7: Performance Analysis

6. CONCLUSION

This study has shown that machine learning is an effective technique in identifying fake job advert. The proposed system has combined NLP and classification methods and presents an interface as a web-based system. The system can drastically eliminate the possibility of job fraud, with an accuracy rate of more than 95%. Future research will be centred on:

- Supplementing the data with real- time job posting.
- Applying deep learning models (e.g. LSTM, BERT) for more semantic deep learning understanding.
- Making the system available as an API to be integrated into significant job sites as a cloud-based.

7. REFERENCES

[1] Gupta M. K., Sharma R., Job Scam Detection using NLP and Ensemble Models, https://ieeexplore.ieee.org/abstract/document/9569182, 2021.

[2] Zhang Y., Fake job post detection by machine learning, Journal of AI Research, 2020.

[3] Chugh aware G., Arora P., Linguistic Features to Scam Job Detection Springer LNCS, 2022.

[4] The ILO Publications report titled, World Employment Report, 2023, Online Recruitment Fraud.

[5] Kaggle Dataset: “Fake Job Postings Dataset,”[Available at:] Accessible through the link: https://www.kaggle.com/datasets/shivamb/real-or-fake-job-postings