A SYNOPSIS ON

Job Fraud Prediction

SUBMITTED

BY

Saikat Maity – TNU2022053100038L

Anas Rahaman Molla - TNU2021053100033

Shovan Bera – TNU2021053100030

Arindam Pakhira - TNU2021053100027

Ritaban Guchait - TNU2021053100020

FOR

THE PARTIAL FULFILLMENT OF B. Tech DEGREE

IN

CSE (AI & ML)

UNDER THE GUIDANCE

OF

<< Dr. Usha Rani Gogoi >>

2021-2025

SCHOOL OF SCIENCE AND TECHNOLOGY

THE NEOTIA UNIVERSITY

INTRODUCTION:

⬤ **Objective**:

- To develop a machine learning-based system to predict and identify fraudulent job postings.

- Enhance job seekers' safety by reducing exposure to potential scams.

⬤ **Dataset**:

- Use publicly available datasets from job portals or synthetic datasets designed for job fraud detection.

- Features may include job title, description, salary range, company information, and metadata like email domains and links.

⬤ **Key Techniques**:

- Natural Language Processing (NLP) for analyzing textual data such as job descriptions.

- Supervised learning algorithms like Random Forests, or Neural Networks for classification.

PROBLEM STATEMENT:

⬤ **Imbalanced Datasets**:

- Job fraud detection suffers from highly imbalanced datasets, where fraudulent job postings are significantly fewer than legitimate ones. This imbalance makes it challenging for machine learning models to effectively identify fraud without the risk of bias toward legitimate posts.

⬤ **Lack of Contextual and Textual Analysis**:

- Many traditional fraud detection methods rely on structured metadata alone, ignoring the rich, unstructured data found in job descriptions and other textual content. This limits the ability to identify nuanced fraud patterns hidden within the text.

⬛ **Need for Real-Time Fraud Detection**:

- There is a growing need for an automated, real-time fraud detection system that can analyse job postings instantly and accurately flag potentially fraudulent jobs before they are viewed by users.

⬛ **Scalability and Adaptability**:

- Existing fraud detection systems may struggle to scale across large platforms with millions of job postings or adapt to new fraud tactics, making them less effective over time as fraud evolves.

OBJECTIVE AND SCOPE:

Objective:-

⬛ **Develop a Machine Learning Model**:

- Build a predictive machine learning model to automatically classify job postings as fraudulent or legitimate based on multiple features (textual, metadata, etc.).

⬛ **Enhance Job Seeker Safety**:

- Provide job seekers with an automated tool that flags potentially fraudulent job postings in real-time, reducing exposure to scams and protecting personal and financial information.

Scope:-

⬛ **Dataset and Feature Scope**:

- The project will focus on job postings from publicly available job portals, incorporating textual content (job descriptions), metadata (company name, location, salary), and metadata from email domains and job posting URLs.

⬛ **Fraud Types Scope**:

- The project will focus on detecting common types of job fraud, including fake job postings, phishing scams, and posts that request personal information or payment from applicants.

METHODOLOGY:

Data Collection:

- Gather data from publicly available job portals or synthetic datasets.
- Include features like job title, description, salary, company name, email domains, and other metadata.

Data Preprocessing:

- Handle missing or incomplete data through imputation or removal.
- Perform text preprocessing for job descriptions (tokenization, stop word removal, stemming/lemmatization).
- Normalize numerical features and encode categorical variables using techniques like one-hot encoding.

Exploratory Data Analysis :

- Analyse the dataset to understand feature distributions and identify correlations.
- Visualize data to detect anomalies and understand patterns in fraudulent postings.
- Address class imbalance using techniques like SMOTE (Synthetic Minority Oversampling Technique).

Model Development:

- Split data into training and testing sets to evaluate model performance.
- Use hyperparameter tuning (Grid Search or Random Search) to optimize model performance.

Model Evaluation:

- Evaluate the model using metrics such as precision, recall, F1-score, and accuracy.
- Special attention to recall (minimizing false negatives) to ensure fraudulent cases are not missed.

HARDWARE & SOFTWARE:

Google Collab – Platform to make the project and experiment on model

Python Libraries – NumPy, Pandas, TensorFlow

ML Model – Random Forest, Vectorization

DL Model – RNN Model, LSTM Model, GRU Model.

The Project needs to be approved by the respective Guide :

Dr. Usha Rani Gogoi

FUTURE PLAN OF ACTION:

✓ **Collaboration with Industry:**

• Partner with job portals and recruitment agencies to improve dataset quality and integrate the model into real-world systems**.**

✓ **Integration with Cybersecurity Tools:**

• Combine job fraud detection with broader cybersecurity measures, such as phishing detection, for holistic protection.

.