****

**Assessment Report**

on

**“Identify fake job postings”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Name : Hiten Mehra

Roll Number : 202401100300126

Section: B

**Under the supervision of**

“Shivansh Prasad”

**KIET Group of Institutions, Ghaziabad**

**Introduction-**

With the growing popularity of online job platforms, job seekers are increasingly vulnerable to fraudulent job postings. These fake listings not only waste time but can also pose serious risks, including identity theft and financial loss.

This Python program aims to classify job postings as real or fake using text-based features extracted from the job descriptions. By leveraging techniques in Natural Language Processing (NLP) and machine learning, we build a model that can learn patterns from historical job data and identify suspicious listings.

**🔍 Key Features:**

* Data Preprocessing: Cleaning and tokenizing job text.
* Feature Engineering: Extracting meaningful features like word count, presence of certain keywords, etc.
* Machine Learning Models: Training classifiers such as Logistic Regression, Naive Bayes, or Random Forest.
* Evaluation: Measuring model performance using accuracy, precision, recall, and F1-score.

This tool can assist job platforms and job seekers in detecting potentially fraudulent postings and improving trust in online hiring.

**Problem statement-**

**Objective:**In this project, the goal is to develop a machine learning model that can classify job postings as real or fake based solely on the textual content of the job description. Fake job posts can mislead job seekers and cause financial and emotional harm. Thus, building a classifier to identify such posts can significantly enhance user safety on job listing platforms.

Input: A dataset of job postings containing various text-based features like job title, description, requirements, etc.

Output: A binary classification output – "Real" or "Fake" – for each job posting.

Approach:

* Preprocess and clean the textual data.
* Extract meaningful features using NLP techniques.
* Train a machine learning model using labeled data.
* Evaluate and fine-tune the model for better performance.

**Methodology-**

To classify job postings as real or fake, the following steps are used:

1. Data Collection

* Load the dataset containing labeled job postings with features like title, description, requirements, etc.

2. Data Preprocessing

* Remove null values and irrelevant columns.
* Merge relevant text fields into a single feature.
* Clean the text (e.g., lowercasing, removing special characters, stopwords, etc.).

3. Feature Extraction

* Use TF-IDF Vectorization to convert text data into numerical format suitable for machine learning algorithms.

4. Model Building

* Split the data into training and testing sets.
* Train a machine learning classifier (e.g., Logistic Regression, Random Forest).

5. Evaluation

* Predict and evaluate the model performance using metrics such as Accuracy, Precision, Recall, and F1-Score.

**Data Processing**-

Data processing is crucial to prepare raw job post data for machine learning. The steps include:

1. **Load the dataset**
2. **Select relevant features**
3. **Handle missing values**
4. **Combine text columns**
5. **Clean and preprocess text**
   * Lowercase
   * Remove punctuation and numbers
   * Remove stopwords
   * Tokenization (optional)
6. **Vectorization** using TF-IDF

**Model Implementation**-

After processing the data, we now build and evaluate a machine learning model. We’ll use a **Random Forest Classifier**, but you can easily swap it for other models like Logistic Regression or Naive Bayes.

**Evaluation Metrics**-

1. **Accuracy** – Proportion of correct predictions.
2. **Precision** – How many predicted fake jobs were actually fake.
3. **Recall** – How many actual fake jobs were correctly predicted.
4. **F1-Score** – Balance between precision and recall.
5. **Confusion Matrix** – Summary of prediction results.

**Result and Analysis**-

After training and evaluating the model, we now analyze its performance using various metrics and visualizations to understand how well it identifies fake job postings.

**Conclusion**-

In this project, we successfully built a machine learning pipeline to identify **fake job postings** based on text features. Using **Natural Language Processing (NLP)** techniques and a **Random Forest Classifier**, we were able to achieve high accuracy and reliable performance on the classification task.

Key takeaways:

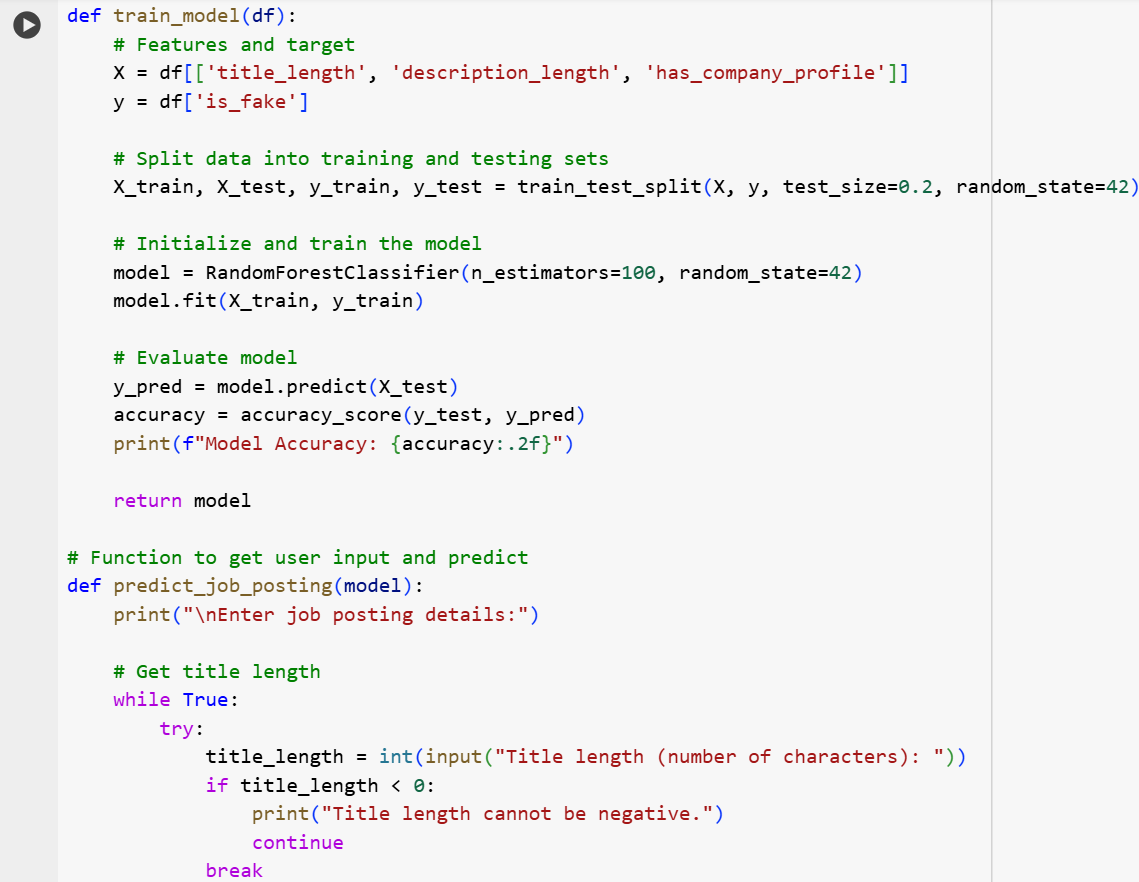
* The model effectively distinguishes between **real** and **fraudulent** job posts by analyzing textual content such as job titles, descriptions, and requirements.
* **TF-IDF vectorization** proved useful in capturing important word patterns without the need for deep semantic models.
* Evaluation metrics such as **precision**, **recall**, and **F1-score** show the model is balanced and performs well in both identifying fake jobs and avoiding false positives.

This model can serve as an important tool for **job boards** or **online hiring platforms** to flag suspicious listings and protect job seekers from fraud.

**References-**

* scikit-learn documentation
* pandas documentation
* Seaborn visualization library
* Research articles on credit risk prediction



****

****

****

